BELIEVABLE SUSPECT AGENTS

RESPONSE AND INTERPERSONAL STYLE SELECTION FOR AN ARTIFICIAL SUSPECT

MERIJN BRUIJNES

BELIEVABLE SUSPECT AGENTS

Response and Interpersonal Style Selection
for an Artificial Suspect

Merijn Bruijnes

Ph.D. Dissertation Committee:

Chairman and Secretary:

Prof. dr. P.M.G. Apers	University of Twente, NL

Supervisor:

Prof. dr. D.K.J. Heylen	University of Twente, NL

Co-Supervisor:

Dr. H.J.A. op den Akker	University of Twente, NL

Members:

Prof. dr. C. Pelachaud	TELECOM ParisTech, FR
Prof. dr. M.G.J. Swerts	Tilburg University, NL
Dr. T. Bosse	Vrije Universiteit Amsterdam, NL
Prof. dr. E. Giebels	University of Twente, NL
Prof. dr. V. Evers	University of Twente, NL

Paranymphs:

Brendan Bruijnes

Florian Bruijnes

CTIT Ph.D. Thesis Series ISSN: 1381-3617, No. 16-408
Center for Telematics and Information Technology
P.O. Box 217, 7500 AE Enschede,
The Netherlands

SIKS Dissertation Series No. 2016-39
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

The research reported in this dissertation was supported by the Dutch national program COMMIT.

The research reported in this dissertation was carried out at the Human Media Interaction group of the University of Twente.

Cover design by Loes van Splunter. Typeset with LaTeX. Printed by Netzodruk

ISBN: 978-90-365-4203-6
DOI: 10.3990/1.9789036542036

BELIEVABLE SUSPECT AGENTS

RESPONSE AND INTERPERSONAL STYLE SELECTION
FOR AN ARTIFICIAL SUSPECT

DISSERTATION

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus
Prof.dr. H. Brinksma
on account of the decision of the graduation committee,
to be publicly defended
on *Friday October 7, 2016*, at *16:30*.

by

Merijn Bruijnes
Born October 9, 1984
in Oldebroek, the Netherlands

This dissertation has been approved by:
Supervisor: prof. dr. D.K.J. Heylen
Co-Supervisor: dr. H.J.A. op den Akker

Contents

Acknowledgements

Here I want to thank everyone who, in some way or another, shared in –and contributed to– my PhD track. It was an awesome time thanks to all of you! But let me start at the beginning. I first learned about the Human Media Interaction (HMI) group through Gijs, who started a PhD at the HMI group when we still were housemates. Someday, Gijs told me that someone from HMI, Rieks, might be looking for a PhD student. That seemed like a nice opportunity and I wrote an e-mail to see if he would consider me for that position. Just as I had sent it, Gijs told me that a large project, a COMMIT project, was awarded to HMI. So I wrote another e-mail to the chair of HMI at that time, Anton, to congratulate him and ask if this meant that there would be a PhD position available and if he would consider me. I received an invitation from both Rieks and Anton (to this day I don't know whether they discussed this beforehand). The best thing is, they gave me the opportunity to pursue this dream with Rieks as my co-supervisor and Dirk as my supervisor. For this I will remain forever grateful.

Rieks, without your help I would not have reached this point. Besides being the best supervisor someone could wish for, I also really enjoyed our discussions about anything, everything, nothing, and whether or not something else might be part of anything. Thank you for always being there, for letting me figure out want I wanted to investigate, for having the patience to deal with me, and for always being enthusiastic.

Dirk, I really appreciate that, despite that I was hired by Anton, you accepted me as one of your own PhD students. I came to enjoy our discussions and the difficult questions that you always asked. I feel you helped me grow up. And of course, thank you for the wonderful extracurricular activities such as molecular cooking, Thunder Thursdays, sticking the head of a German guy on a French guy's body, sponsoring our brewery, and signing off on all of our travels.

The people at HMI are awesome! Always ready to go, whether it is to participate in an experiment, talk statistics, try to teach me programming, playing *Fussball!*, having lunch breaks, or just hanging out [198]. I want to mention my roommates over the years: Alejandro M., Bart, Danish, Dong, Jelte, Robby, Thijs en Thomas. You guys (and girl) rock! You are what made our side of the hallway 'the cool side'. And even if you weren't here to make our office the best office, we were lucky to always have *the biggest office*! But all of my colleagues at HMI are awesome: the lunch group A (we won, there is no more lunch group B: just a bigger and better lunch group A!), the people who consistently beat me at Fussball (I will get you next time) and the people who were kind enough to let me win, the eNTERFACE'13 group who were brave enough to make the daily trips between down town Lisbon and the Universidade

Nova de Lisboa with me as their regular driver, all my COMMIT-buddies, and last but not least the HMI staff. Thanks Jeroen, Merel, Christian, Aduen, Christina, Randy, Michiel, Jered, Bob, Jaebok, Khiet, Daniel, Daphne, Alejandro C., Dennis, Mariët, Betsy, Mannes, Jan v. E., Gwenn, Angelika.

A special thanks for all the support from 'my mothers at work': Lynn, Charlotte, Alice, and Wies. Lynn, thank you for reading and correcting all my spelling and grammatical mistakes and then having the patience to explain to me what was the difference between English and 'MerijnEnglish'. Any mistakes in the current version of this thesis are obviously mistakes that I re-inserted myself!

I always enjoyed working together with students. Some of them really helped me in my research and I even published together with some (as such those students are already mentioned in the bibliography). Special thanks to Jan, Sophie, Merijn S., Sjoerd, Rens, Rifca, and all other students. Also, I should mention the 'Explore the void' study trip organised by Creative Technology students. The trip to Silicon Valley was life changing and I had the pleasure to tag along as supervisor together with Erik.

I want to thank the people from COMMIT, they were more than 'people who gave the money' but turned out to be 'people who created a community'. The help of the Dutch police and the Dutch police academy was crucial to this thesis. In particular I want to thank (in no particular order) Ron, Hans, Arend, Ronald, Siemen, Machteld, Roel, Bert, Manon, Esther, Willem and all the police academy students for their efforts.

I want to thank my family and friends. I am very lucky that my family are also my friends. Gerrit, Ina, Brendan, Florian bedankt voor het altijd voor me klaarstaan, voor alle gezelligheid, voor de vele uren naar mijn verhaal luisteren, voor alle grappen, voor het me opzoeken (zelfs in Lissabon) als ik weer eens tegenslag had in de liefde, voor alle vakanties, voor het tot diep in de nacht gamen, voor alle logeerpartijen, voor het samen biertjes drinken en dan de amateurfilosofen uithangen. Zonder jullie had ik het nooit gehaald, maar met jullie kan ik alles aan!

Aan mijn vrienden van lang geleden: Ger-Jan en Pieter, mooi dat we elkaar nog af en toe zien! Aan mijn vrienden van iets minder lang geleden: Arjan, Koos, Menno, Gijs, Lucas, Eva, Sabine, Henk Skatoelakis en Koenoepi. Ik heb een top studententijd gehad dankzij jullie! Aan mijn nieuwste vrienden: Bram, Robby, Ronald, Jan, Alejandro, Jeroen, Frans. Ik hoop dat we elkaar blijven zien na onze promoties! Goed, dat waren de meesten wel. Sorry aan iedereen die ik vergeten ben of niet bij naam genoemd heb (zoals alle aanhang en de bootgroep), jullie horen er natuurlijk wel bij!

Dan de laatste, maar zeker niet de minste persoon die ik wil noemen: mijn vriendin Manon. Bedankt voor al je liefde, je geduld en begrip voor dat je soms alleen naar bed moest omdat ik weer eens een deadline had en de vele keren dat je voor me hebt gekookt. Ik hoop dat we nog vele jaren vele mooie nieuwe avonturen gaan beleven in verre landen, in bijzondere restaurantjes en in onze fantasie, en dat we voor altijd 'meer dan buren' zullen zijn.

And with that, I can take the last item of the TODO-list:

TODO: ~~Write the only chapter that everyone reads.~~

English Summary

The social skills necessary to properly and successfully conduct a police interrogation can and need to be trained. In this thesis I will describe the steps I took towards a virtual character that can play the role of a suspect in a police interrogation training. Students of the police academy will be able to use this 'virtual suspect' to practise their social skills.

The virtual suspect needs to behave as a human suspect would. An important first step towards this goal is an analysis of the behaviour of human suspects in a police interrogation. We collected a corpus of practise police interrogations: the Dutch Police Interview Training (DPIT) corpus. This corpus contains recordings of professional training actors who played the role of suspects who were interrogated by students of the Dutch Police Academy.

Leary's theory on interpersonal stance is used by the police as a theoretical framework to understand the social dynamics in a police interview. Using the concepts dominance and affiliation, the theory describes how suspects take stance during an interview and how this is related to the stance that the interviewer takes. In chapter 3, we describe whether observers could agree on what interpersonal stance was taken in the DPIT corpus. It turned out that agreement between observers was very low on the level of individual turns of speech. However, we showed that a 'majority vote' of multiple observers can indeed reveal the dynamics of stance taking in the entire interview. The agreement of individual observers with this majority vote was higher than between individual observers. Subsequent analyses of the disagreement of observers revealed that this 'noise' was not random, but that for one speech turn the selected labels of all observers often center around one stance. A 'fuzzy' noise filter applied to simulated judgements of stance showed similar results. Agreement between simulated observers was very low, yet a majority vote of multiple simulated observers did reveal the dynamics of stance taking in the entire interview. Also, the majority vote of the simulated observers showed the dynamics of the interview that were used as input for the simulation. From this we concluded that although inter-annotator agreement on stance labelling on the level of speech segments is low, external observers are able to reveal the important dynamics in stance taking in a police interview.

Next, we explored the relation between the stance taken by the suspect and the turn-taking behaviour. Turn-taking behaviour consists of overlaps, interruptions, pauses, and silences in a conversation. This explorative study into the relation between a suspect's stance and the types of overlaps, interruptions, and silences indicated that the interview topic and in particular how the topic was related to the case at hand

was an important factor that influenced the stances taken by the subject. Stances and roles seemed to be mediating factors for the meaning of overlaps and silences in suspect interviews.

Observers did not show high agreement when asked which stance a person showed. Therefore we investigated whether human judges agreed on the way they perceived the various aspects of stance taking. Eight amateur actors acted as a suspect in a police interrogation with four different stances. These recordings were shown in an online survey to participants who described them by selecting a number of adjectives from a list. We computed the inter-rater agreement with Krippendorff's alpha statistics using a distance metric that was based on the theory of interpersonal stance. Results showed that for some of the stance types observers agreed more than for others.

We analyse the behaviour of actors and not real suspects to investigate how a virtual suspect should behave in order to behave as a human suspect. Therefore it is important to investigate the effect of using actors. We did this by investigating whether the proficiency of the actor had an effect on the reliability with which observers interpreted the actor's actions. We compared the fragments from amateur actors to fragments from professional actors taken from popular TV-shows. We found that some actors are better at portraying an interpersonal stance than others. Also, validity (recognizing which stance is acted) and agreement between observers did not always go hand in hand.

The virtual suspect should behave as a human suspect would behave and thus it should respond to the human interrogator as a human suspect would respond. For this we need to understand *why* a suspect responds in the manner he or she responds. Therefore, the next step is investigating which social and psychological theories can give an explanation for the behaviour of a (human) suspect in a police interrogation. These theories can then be used to create a model that can determine appropriate behaviour of the virtual suspect in response to the behaviour of the interviewer. We analysed the DPIT corpus to get insight into which social behaviours occur in this setting. We collected the terms observers used to describe the interactions in the interviews. Through factor analysis of this 'semantic model', we showed that the theories interpersonal stance, face, and rapport and the meta-concepts information and strategy are necessary to include in a model that captures the social interaction in an interrogation. Subsequent validation and relational analysis of the concepts from these theories showed which concepts from these theories were related. We used these theories to create a model that can determine an appropriate response to the social action of an interrogator: a Response Model.

A virtual agent needs three main capabilities to be able to have a meaningful social interaction with a user. The capability to 'sense, think, and act'. As well as this, we distinguish *what* is being said and *how* it is being said. Our Response Model primarily deals with *how* something is being said. The actions of the user have to be sensed and interpreted, next it should be reasoned what response is appropriate to this action and finally this response should be acted out. The dialogue action of a user, for example "*Tell me when you were there, crook!*" can be interpreted in abstract terms, for example as a dominant and aggressive solicitation of information. This interpretation can be used to reason what response is appropriate. It could be reasoned an aggressive re-

sponse without the requested information is appropriate, for example "*Go figure it out yourself!*". Our Response Model uses an abstract representation of the social action of the user, and the personality of the virtual suspect, to determine the interpersonal state in the conversation. This information is used to determine an abstract representation of the social reaction of the virtual suspect. This abstract response contains for instance how friendly the answer should be and whether or not the truth should be told.

The credibility of virtual humans, such as the virtual suspect, is crucial for a 'serious game' with which users can train their social skills. Users need to be willing to join in the role-play and the probability that they will do this is greater when the game has a compelling story and realistic virtual characters. This requires consistency within the possible behaviours of the virtual human: the virtual human should behave in a manner that is in agreement with his or her nature. The Response Model contains a number of personality settings which can describe the personality of the virtual suspect. This allows the system to play different suspects depending on what the persona of a suspect is according to the scenario of the game. To what extent the virtual human can behave as a human is a consequence of the components that make up the system. Each component has a specific task and performs this task with some level of human-likeness. We evaluated the Response Model in two variants: separate from other components and integrated in a complete virtual suspect system. For the evaluation we defined three different personality settings and we described these in personas. The evaluation had the form of a 'Guess who you are talking to?' task. Participants interacted with the virtual suspect and were unaware of its personality setting. Afterwards the participants were asked to choose which of a number of personas was most similar to the personality of the suspect they had just interacted with. In the variant where participants interacted with the Response Model separated from other components, they used abstract descriptions of 'how-I-would-say-it'. For example, the participant would input 'friendly' and 'statement' to which the response model responded with 'cooperative' and 'question'. In this manner participants were able to interact with the isolated Response Model. In the variant where the Response Model was integrated with the complete virtual suspect system, participants could use free speech to interact. The embodied virtual suspect responded to *what* they said and *how* they said it. Participants were able to 'Guess who they were talking to' better than chance in both variants. In the integrated variant participants were less often able to correctly guess which personality setting they had interacted with than in the separated variant. Additionally, we found that personas that differed more were less likely to be confused. This means the response model was indeed able to select different behaviour for different personas and that the behaviour differed more when the personas were more different. Finally, we argued that some participants managed to change a persona's initial mood and overcome its personality so that it showed behaviour not characteristic for the persona.

Nederlandse Samenvatting

Sociale vaardigheden kunnen geoefend worden. In sommige beroepen zijn goede sociale vaardigheden cruciaal bij het uitvoeren van bijbehorende taken. Bij de politie moet ondervragingsvaardigheid, een taak die goede sociale vaardigheden vereist, geoefend worden. In dit proefschrift beschrijf ik de stappen die we gezet hebben om een virtueel karakter te creëren dat de rol van een verdachte in een oefenpolitieverhoor kan spelen. Studenten van de politieacademie kunnen in de toekomst hun sociale verhoorvaardigheden oefenen met deze 'virtuele verdachte' in een rollenspel.

De virtuele verdachte moet zich gedragen zoals een menselijke verdachte zich zou gedragen. Een belangrijke eerste stap hiertoe is een analyse van het gedrag van menselijke verdachten in politieverhoren. Wij verzamelden opnamen van (oefen-) politieverhoren: het Dutch Police Interview Training (DPIT) corpus. Op de beelden in dit corpus is te zien hoe professionele trainingsacteurs worden verhoord door studenten van de politieacademie.

Een belangrijke theorie over de sociale dynamiek in een verhoorsituatie die door de politie wordt gebruikt is de interpersonal stance theorie van Leary, ook wel Leary's Rose genoemd. Deze theorie beschrijft, in termen van dominantie en vriendelijkheid, hoe de verdachte en de verhoorder elkaar beïnvloeden met hun gedrag en gesprekshouding. Een gesprekshouding die iemand aanneemt bestaat uit hoe dominant en vriendelijk iemand is richting de ander en wordt 'stance' genoemd. In de eerste studie, in hoofdstuk 3, onderzochten we of het mogelijk was voor observatoren om overeenstemming te bereiken over welke stances zij dachten dat er ingenomen werden bij elke spreekbeurt door de verhoorder en de verdachte. Dit bleek lastig: de overeenstemming tussen verschillende beoordelaars was erg laag. Toch bleek het mogelijk om een hogere overeenstemming te bereiken door te kijken naar welke stance de meeste stemmen kreeg: een 'majority vote'. De overeenstemming van individuele beoordelaars met deze majority vote was hoger dan tussen individuele beoordelaars. Deze majority vote onthulde de dynamiek van stances in het gehele verhoor waar de meeste beoordelaars zich in konden vinden. Uit verdere analyse bleek dat de verschillen tussen beoordelaars, de ruis, voor een groot deel niet op toevalligheid berustte, maar dat de gerapporteerde stances van een bepaalde gespreksbeurt zich vaak rond één stance centreerden. Gesimuleerde beoordelingen van stance die voorzien werden van een 'fuzzy' ruis lieten een vergelijkbaar beeld zien qua overeenstemming. Ook hierbij was er sprake van lage interbeoordelaarsovereenstemming, maar de overeenstemming met een majority vote was een stuk hoger. The majority vote van de gesimuleerde observatoren liet de dynamiek zien zoals deze als input aan de simulatie

werd gegeven. Dit bevestigde het idee dat herkenning van stance door externe observatoren mogelijk is en zinnige resultaten oplevert ondanks lage overeenstemming tussen individuele observatoren.

Vervolgens onderzochten we de relatie tussen de stance van een verdachte en het beurtwisselingsgedrag in een gesprek. Overlappende spraak, in de rede vallen, pauzes en stiltes zijn voorbeelden van beurtwisselingsgedrag. We vonden dat het besproken onderwerp van grote invloed is op de stance die een verdachte aanneemt. Dit is vooral het geval bij onderwerpen die met de verdenking te maken hebben. De stance, maar ook de rol van de spreker, beïnvloedt de betekenis van overlappende spraak en stiltes.

Observatoren bleken geen hoge overeenstemming te hebben over welke stance iemand liet zien. Daarom onderzochten we of observatoren op dezelfde manier de verschillende aspecten van het aannemen van een stance waarnemen. Amateuracteurs acteerden met vier verschillende stances als een verdachte in een verhoor. Deze opnames werden in een online-enquête aan proefpersonen getoond. Proefpersonen beschreven de getoonde stance door middel van een selectie uit een lijst van adjectieven. We gebruikten Krippendorff's alpha met een afstandsmetriek gebaseerd op theoretische verklaring als statistische methode om een maat voor de overeenstemming tussen de verschillende lijsten met adjectieven te berekenen. Uit de resultaten bleek dat observatoren het vaker eens zijn bij sommige geacteerde stances dan bij andere geacteerde stances.

Voor onze analyses keken wij naar het gedrag van acteurs die een verdachte spelen, en niet naar echte verdachten. Hierdoor is het belangrijk te bepalen wat het effect is van het gebruik van acteurs. Wij deden dit door te onderzoeken of de bekwaamheid van een acteur invloed heeft op de betrouwbaarheid waarmee een observator de acties van een acteur interpreteert. Het bleek dat sommige acteurs beter zijn in het laten zien van een stance dan andere acteurs. Bij het beoordelen van welke stance iemand laat zien gaan validiteit (herkennen welke stance geacteerd werd) en overeenstemming tussen waarnemers niet altijd hand in hand.

De virtuele verdachte moet zich gedragen zoals een menselijke verdachte zich zou gedragen en dus reageren op de menselijke verhoorder zoals een menselijke verdachte zou reageren. Het is nodig te weten *waarom* een verdachte reageert zoals hij of zij reageert. De volgende stap is daarom onderzoeken welke psychologische en sociale theorieën een verklaring kunnen geven voor het gedrag van een (menselijke) verdachte in een politieverhoor. Deze theorieën kunnen dan gebruikt worden om een model op te stellen waarmee een virtuele verdachte op basis van gedrag van de ondervrager een reactie kan bepalen. Hiertoe analyseerden wij het DPIT corpus om inzicht te krijgen welke sociale gedragingen in een politieverhoor voorkomen. We verzamelden de termen die observatoren gebruiken om de interacties in het verhoor te beschrijven. Door factor analyse toe te passen op dit 'semantische model' verkregen wij clusteringen van termen welke geïnterpreteerd konden worden als concepten die bestaan binnen psychologische en sociale theorieën. Het bleek dat volgende theorieën van belang zijn in een verhoor: interpersonal stance (interpersoonlijke houding), face (gezicht als in gezichtsverlies), rapport (sociale band) en de meta-concepten informatie en strategie. Deze theorieën hebben wij vervolgens gebruikt om een model op te stellen wat op een sociale actie van een verhoorder de sociale reactie van een virtuele

verdachte kan bepalen: het Response Model.

Een virtuele verdachte heeft drie cruciale capaciteiten nodig om een zinvolle sociale interactie te kunnen hebben. De capaciteit om te kunnen waarnemen, redeneren en acteren, ook wel 'sense, think, act' genoemd. Verder onderscheiden wij *wat* iemand zegt en *hoe* iets wordt gezegd. Ons Response Model houdt zich primair bezig met *hoe* iets gezegd wordt. Een actie van de gebruiker moet worden gedetecteerd en geïntepreteerd, vervolgens moet beredeneerd worden welke reactie op deze actie gepast is en daarna moet deze reactie uitgevoerd worden. De dialoogactie van een gebruiker, bijvoorbeeld "*Zeg dat je het gedaan hebt, crimineeltje!*" kan worden geïnterpreteerd als een agressief en dominant verzoek om informatie. Deze interpretatie kan gebruikt worden om te beredeneren welke reactie gepast is. Zo zou geredeneerd kunnen worden dat een agressieve dominante uitspraak zonder de gevraagde informatie gepast is. Het abstracte resultaat van de beredenering kan omgezet worden in een concrete actie die voldoet aan de abstracte reactie, bijvoorbeeld "*Zoek het uit, bromsnor!*". Ons Response Model gebruikt een abstracte representatie van de sociale actie van de gebruiker en de persoonlijkheid van de virtuele verdachte om de interpersoonlijke staat te bepalen. Deze informatie wordt vervolgens gebruikt om een abstracte representatie van een sociale reactie van de virtuele verdachte te bepalen. Deze representatie bestaat uit onder andere hoe vriendelijk de reactie zal worden en of de waarheid gesproken zal worden. Deze informatie wordt gebruikt om het gedrag van de virtuele verdachte aan te sturen.

De geloofwaardigheid van virtuele mensen, zoals de virtuele verdachte, is cruciaal voor een 'serious game' waarmee gebruikers sociale vaardigheden trainen. Gebruikers moeten mee willen gaan in het rollenspel en de kans dat zij dit doen is groter als er sprake is van een meeslepend verhaal en realistische virtuele karakters en gebeurtenissen. Hiervoor is consistentie binnen de mogelijke gedragingen van het virtuele karakter nodig: een virtueel karakter moet zich gedragen op een manier die in overeenstemming is met zijn of haar natuur. Het Response Model bevat een aantal persoonlijkheidsinstellingen welke de persoonlijkheid van de virtuele verdachte kunnen beschrijven. Op deze manier kan het systeem verschillende verdachten spelen al naar gelang wat voor een verdachte het scenario vereist. In hoeverre een virtueel mens zich als een echt mens kan gedragen is een een gevolg van de mogelijkheden van de verschillende componenten waaruit het systeem bestaat. Elk van de componenten heeft een specifieke taak en voert deze taak uit met een bepaalde mate van menselijkheid. Tijdens de evaluatie van het Response Model onderscheidden we daarom twee varianten van het systeem: het Response Model geïsoleerd van andere systemen, en het Response Model geïntegreerd in een volledig functionele virtuele verdachte. Voor de evaluatie werden drie persoonlijkheidsinstellingen bepaald en beschreven in de vorm van drie personas. De evaluatie had de vorm van een 'Raad met wie u gesproken heeft'-taak. Proefpersonen interacteerden met de virtuele verdachte waarvan zij vantevoren niet wisten hoe de persoonlijkheid was ingesteld. Achteraf kregen de proefpersonen de taak te kiezen welke van een aantal beschreven personas het meest leek op de persoonlijkheid van de virtuele verdachte waarmee zij zojuist hadden geïnteracteerd. In de variant waar proefpersonen interacteerden met het Response Model geïsoleerd van andere systemen, gebruikte men een tekstinterface om

een abstracte beschrijving van een sociale uiting te geven. Zo kon een gebruiker bijvoorbeeld aangeven dat hij of zij een vriendelijke vraag stelde, waarop de virtuele verdachte aangaf dat hij een vriendelijk kort antwoord gaf. Op deze manier konden proefpersonen interacteren met enkel de Response Model compontent. In de variant waarbij het Response Model geïntegreerd was in een volledig functionele virtuele verdachte konden proefpersonen vrije spraak gebruiken om te interacteren. Hierbij reageerde de virtuele verdachte zowel op wat men zei als hoe men het zei. Het bleek dat proefpersonen beter dan kans konden aangeven welke persoonlijkheidsinstelling de virtuele verdachte had in beide varianten. In de geïntegreerde variant wisten proefpersonen minder vaak correct aan te geven met welke persoonlijkheid zij hadden geïnteracteerd dan in de geïsoleerde variant. Verder bleek dat persoonlijkheden die minder op elkaar leken, minder snel verward werden dan persoonlijkheden die meer op elkaar leken. Hieruit concluderen wij dat het Response Model in staat is gedrag te selecteren wat past bij de ingestelde persoonlijkheid van een verdachte.

Part I

Introduction

1

Social Skill Training for Police Interrogations

The social skills necessary to conduct a police interrogation properly and successfully can and need to be trained. In this thesis I will describe the steps I took toward a virtual character that can play the role of a suspect in a practice police interrogation. In this chapter I will explain what a police interrogation is, why training is needed, what this training looks like at the moment, and what it might look like in the future.

1.1 Introduction

A virtual agent that can play a suspect in a serious game that can be used by police students to hone their skills in police interviewing was the goal we worked towards. In this thesis you can read that we came a long way, that we did many studies, and that we made some interesting findings. The question that I always ask myself is: "Why do people do and say what they do and say?". In this thesis I will limit the scope of this question. I have investigated why suspects in a police interrogation say what they say and how they say it. When we have figured this out, we can build a model that can make a virtual suspect respond appropriately when it is being interrogated by a student of the police academy. In part II of this thesis, we will look at the behaviour of suspects that are being interrogated. What kind of behaviour do they show? Can observers agree on what behaviour suspects show? What is the relation between the behaviour of the interrogator and the suspect? These are some of the questions we will ask in part II of this thesis. In part III, I will describe our response model for a virtual suspect and the evaluations of this model. But first, in this chapter, I will explain more about the fascinating world of police interrogations, why it is important to train police students to conduct interrogations, and our vision of how virtual suspects can help train students.

1.1.1 Terminology

Is it 'police interrogation' or 'police interview'? An interview is a non-accusatory question-answer session with a suspect, victim, or witness, whereas an interrogation is intended to elicit the truth from a suspect. In Europe the term 'police interview' is sometimes preferred while interrogation of a suspect is meant. This is because interview has a less forceful connotation than interrogation. In this thesis we will use interrogation and interview interchangeably. We work with the Dutch police so with both words we mean the European less forceful practice of police interviewing.

Another term that is used and that warrants explanation is 'truth'. What is the 'truth' in a police investigation and in an interrogation? It is easy to slip into apparent paradoxes that surround truths and lies, for example '*a person that believes A to be true, tells B is true, while B is in fact true, is not lying*'. The intent of the speaker matters, the speaker must intend to deceive the audience into mistaking what he truly believes, thus leaving them vulnerable to false belief formation. Also, when playing a part in a play, the truth is what the script intends to be true[1]. In this thesis I will take a pragmatic epistemic view and mean that a statement is truthful if it is in accordance with other (forensic) evidence, regardless whether this evidence is part of a thought-up script in a role-play.

In this thesis, I will speak a lot about 'virtual humans'. With this I mean the illusion that is created of a human being in image and voice using computer-generated imagery and sound. Sometimes I will use different terms, for example virtual agent, artificial character, or embodied conversational agent, but unless specified explicitly otherwise, I mean 'virtual human'. One addendum to this is the term 'virtual suspect', which I use to describe a virtual human that has the role of suspect in a role-play.

1.2 Police Interrogations

In this section I will discuss what happens in a police interrogation, and in particular in the Netherlands. A person has to be a suspect before the police is allowed to interrogate him or her. A suspect can be asked to (voluntarily) come in for an interrogation or can be brought in by the police (who can use force if necessary). The Dutch police has the right to detain a suspect for 6 hours (which can be extended to 21 hours). A suspect is usually detained at the police station. During this time the police has the right to interrogate the suspect about his or her involvement in the suspected criminal act. The suspect has the right to remain silent, to a lawyer, and to medical aid should he or she desire so.

The main objective for the Dutch police during an interrogation is gathering knowledge: *establishing the truth*. Despite what popular TV-shows depict, a confession is not the ultimate goal that a 'good-cop' and a 'bad-cop' try to attain. In fact, until 2004 the word *confession* did not occur in the Dutch Code of Criminal Procedure (see [47]). There are few studies on the rate of confessions or denials of suspects during interrogations in the Netherlands. The Research and Documentation Centre of the Netherlands Ministry of Security and Justice reported that about 70% of sus-

[1] See for example: http://www.philosophyetc.net/2005/10/truth-and-lies.html

pects made a full confession, 11% made a partial confession, and 18% denied (these numbers exclude facts that were not covered in the interrogation) [93]. In fact, focussing on confessions is tempting. A study by Kassin et al. [101] showed what police officers from Canada and the USA self-report about police interrogation practices: "Participants estimated that they were 77% accurate at truth and lie detection, that 81% of suspects waive Miranda rights, that the mean length of interrogation is 1.6 hours, and that they elicit self-incriminating statements from 68% of suspects, 4.78% from innocents" [101, p381]. However, Meissner and Kassin [130] have shown that police investigators are significantly more confident but not more accurate than lay people at distinguishing true and false denials. The same was found to be true for distinguishing true and false confessions [102]. With such over estimation of ability, focussing on confessions is a recipe for false confessions.

In the Netherlands, false confessions in poorly executed interrogations and the often co-occurring miscarriages of justice have received a lot of (media) attention. For example, in the Putten murder case (*Puttense moordzaak*) four suspects were arrested and confessed to the murder of a 23-year-old air hostess. The four were convicted to jail despite that their statements kept changing and differed from one another, there was no forensic evidence, and they later retracted their statements. After an extensive (media) campaign, a crime journalist and a former senior constable were eventually able to get the suspects acquitted [23]. It was argued that the false confessions might be attributed to: "frequent and long interrogations, an excessive relationship of trust between the interrogators and the suspects, shaping, confirmation bias, scenario-based interrogations (what could have happened?), the assumption that the suspects had repressed the murder, guided memory, and confronting the suspects with co-suspect's statements and evidence that was wrongly interpreted" [47, p81]. It was realised that it is imperative a suspect gets a platform to present his or her side of the story and to do this without coercing them to shape their story towards any pre-existing beliefs the interrogator might have. It might very well be that he or she is innocent and an open conversation during the interrogation can give the suspect a chance to reveal (new) information relevant to the case that might lead to the culprit.

Police interrogations are a special type of social encounter, primarily because of the role of authority that the police officer has and the often uncooperative stance that a suspect takes—there may be a conflict between the interaction parties. For most people, being a suspect and being interrogated is a very intimidating experience. This brings some suspects to 'shut down', panic, or go along with the police officer because the suspect thinks "*the officer probably knows everything*". Other suspects may have an innate distrust of the police and treat the interviewer with contempt. A reason for such aggressive behaviour might be the 'one-down effect' from the literature about negotiations [58]. It emerges when one party has a distinct disadvantage, for example because of the role he or she is pushed into as occurs with a suspect in an interrogation. The effect is that the person with 'one-down' is more likely to resort to competitive or aggressive tactics in an effort to '(re)gain the upper hand' [58]. Summarising, the police has to deal with often uncooperative, scared, or aggressive suspects. Despite this, the police are tasked to interrogate and 'get information' from suspects. This task requires them to get suspects to cooperate. This task is not at

all easy and it requires training to conduct successfully. In the next sections I will discuss what interrogation methods the police use to get suspects to provide truthful information.

1.2.1 Dutch Interrogation Practices

The Police Academy in the Netherlands has made progress over the past years to improve interrogation practices by developing a method that serves as a standard blueprint for the interrogation of suspects [47]. Currently, all students at the police academy learn this 'standard interrogation strategy' (standaard verhoorstrategie) [202]. It is aimed in particular at suspects that are willing to testify to some extent but who will not make extensive or truthful statements. Simplified, the standard interrogation strategy works by surrounding the question you want to ask. Surrounding a question means that an officer first asks about possible alternative explanations that the suspect can give to this question. After the question is surrounded, the suspect has almost no other choice than to admit/answer the question. For example, the officer observed the car of the suspect driving last night. The officer wants him to admit he was the one driving his car last night. The officer asks him if he has a car. Lets him describe the car. Asks whether he lends his car to others (let's say he never does). Determine it is not stolen. (etc.) Then the officer asks if he drove the car last night. If the suspect denies, the officer shows the photo of the suspect's car driving last night. If the surrounding worked, the suspect has no other option than to admit or provide an alternative explanation.

Training the proper behaviour for interviews is important for the effectiveness of the interview. For example, Holmberg and Christianson [89] showed that when suspects perceive the police officer's behaviour during the interview as dominant they tend to deny criminal accusations. Alternatively, when suspects perceive the interview as humane and respectful they gain the confidence and mental space required to admit criminal behaviour [89]. Richardson et al. [160] investigated the relation between the verbal mimicry (known as Language Style Matching) in police interviews and the confession to criminal behaviour. They showed that interviews that lead to a confession had a higher rate of the suspect matching the verbal language style of the interviewer than interviews that did not lead to a confession. Further, they suggest that language style matching and mimicry can be employed strategically [160, 162]. In other words, the behaviour of the interviewer is of influence on the outcome of the interview and the interviewer can use his or her behaviour strategically. Therefore it is critical to train and maintain the skill of police officers to behave properly when conducting an interview.

1.3 Training Police Interrogators

Despite the growing importance of forensic research and the implementations of new laws that improve the protection of the suspect, interviews are still one of the most important means employed in crime investigations [88, 89, 135, 177]. This holds for interviewing witnesses and victims as well as for interviewing suspects. Extensive

research has resulted in a broad consensus and agreement about how police interviews should be conducted. Research has shown that intensive training can change interviewing behaviour, but also that benefits are obtained only when extensive efforts are made to enhance the maintenance of the learned interview practices [115]. Interview training, often with actors playing the role of a suspect, is expensive and time consuming. Serious games with virtual suspects have potential in training interview tactics [124]. We study, in co-operation with the Dutch police, how artificial intelligence based on conversational behaviour modelling can enhance police interviewing by building systems that can support the interviewer during the interview or that can be used in training interview skills. The work in this thesis is part of the Dutch COMMIT/ project[2].

1.3.1 Theory about Interrogation

A police interview is often a social situation of conflict. Suspects often do not cooperate with the police officer and the police interview in general, but behave in a confronting manner. Suspects may be withdrawn, defiant or even aggressive towards the police officer. The police officer has the difficult task to get the suspect to cooperate and tell the truth in an interview: resolve the conflict. At the start of a police interview course, Dutch police students receive theoretical training on the use of the theory of interpersonal stance, or as they refer to it 'Leary's Rose' [116] (see section 1.3.1.1). In addition, the use of 'negotiation' strategies is taught. With these a police officer can try to change the behaviour of the suspect. The *Table of 10* by Giebels [66] describes the strategies police officers can use when, for example, they want to convince the suspect that cooperation will be of mutual benefit. Strategies are among others: 'Emotional Appeal', 'Rational Convincing', and 'Direct Pressure' (see Table 1.1). After studying the theory, some students get the opportunity to apply what they have learnt in a role-playing exercise with professional training-actors that play the role of suspect.

1.3.1.1 Interpersonal Stance

The Dutch police relies heavily on the theory of interpersonal stance to characterise the social behaviours that occur in the social interaction of the interview. Interpersonal stance (or attitude) refers to "those spontaneous or strategically employed affective styles that colour interpersonal exchanges" [170]. Compared to personality, attitudes are subject to a greater degree of variation over time. Interpersonal attitudes are essentially an individual's conscious or unconscious judgement of how they feel about and relate to another person while interacting with them. Argyle [4] and Leary [116] identify two fundamental dimensions of interpersonal attitudes that can account for a great variety of nonverbal behaviour: affiliation (ranging from positive to hostile) and power (from dominant to submissive).

In training conversational skills, in particular in training to interview suspects, police trainees in the Netherlands use T. Leary's theory of interpersonal relations as a

[2] http://www.commit-nl.nl/

Table 1.1: Table of 10 by Giebels (translated from Dutch from [202]).

#	*Strategy*	**Principle**	**Description**
1	*Be Nice*	Sympathy	Show willingness to talk, empathize.
2	*Be Equal*	Equality	Emphasize commonalities, name external foes.
3	*Be Credible*	Authority	Show trustworthiness, show expertise.
4	*Emotional Appeal*	Self-Perception	Play on feelings (consider victims), offer to earn respect.
5	*Intimidation*	Insecurity	Warn of consequences, personal attack.
6	*Impose Boundaries*	Scarcity	Deny concessions, ignore opponent.
7	*Direct Pressure*	Repetition	Repeat appeal (plant seed), accomplished fact.
8	*Legitimate*	Legitimacy	Refer to rules and laws, refer to other opinions.
9	*Trade*	Mutuality	Ask for something in return, concession after high commitment.
10	*Convince Rationally*	Consistency	Bring forward arguments, confront with contradictions.

framework for analysing their behaviour towards the interviewee and how the suspect's behaviour is understandable as a reaction to their own behaviour and their style of interviewing, for example, is it more understanding or more offensive. Leary's model, the interpersonal circumplex also known as *Leary's Rose* ([116]) is used in training conversational skills in various professions as for example in health care [184]. Theories similar to Leary's Rose are known under names such as the Interpersonal Checklist [114] and the interpersonal circumplex [164], but the differences are often superficial. The model is presented by a circular ordering of eight categories of interpersonal behaviour, situated in a two-dimensional space spanned by two orthogonal axis. The horizontal axis affiliation (positive versus hostile) the vertical one is the power axis (dominant versus submissive) (Figure 1.1A). The theory says that for the power axis; dominant behaviour is met with submissive behaviour and submissive behaviour is met with dominant behaviour. For the affiliation axis; positive behaviour is met with more positive behaviour and hostile behaviour is met with more hostile behaviour (Figure 1.1B). For example, if someone is dominant and positive towards you (DP, leading or helping), you would gladly cooperate with them which is a submissive and positive (SP) stance. In other words, the level of affiliation between interlocutors is in phase, whereas the level of dominance is out of phase. This phenomenon is also called 'interpersonal complementarity'. However, Orford [144] pointed out that empirical evidence shows a slightly more complicated picture. He showed in a meta-study that friendly-dominant and friendly-submissive behaviour are complementary, but that hostile-dominant behaviour leads to more hostile-dominant behaviour, and hostile-submissive behaviour often leads to dominant-friendly behaviour (see Figure

1.1C). Yet, more recent work by for instance Sadler et al. [166] shows interpersonal complementarity holds up.

It is possible to employ this mechanism strategically. Person A can change his stance in the hope that person B will change his stance according to the theory, thus giving person A some control over person B's stance. One can see why the police is interested in teaching such a theory. They deal with uncooperative suspects and have the task to make them more cooperative. Leary's theory provides a clear strategy to attempt this change in stance of the suspect. For police officers it is an important goal during the interrogation to establish a good working relationship with the suspect.

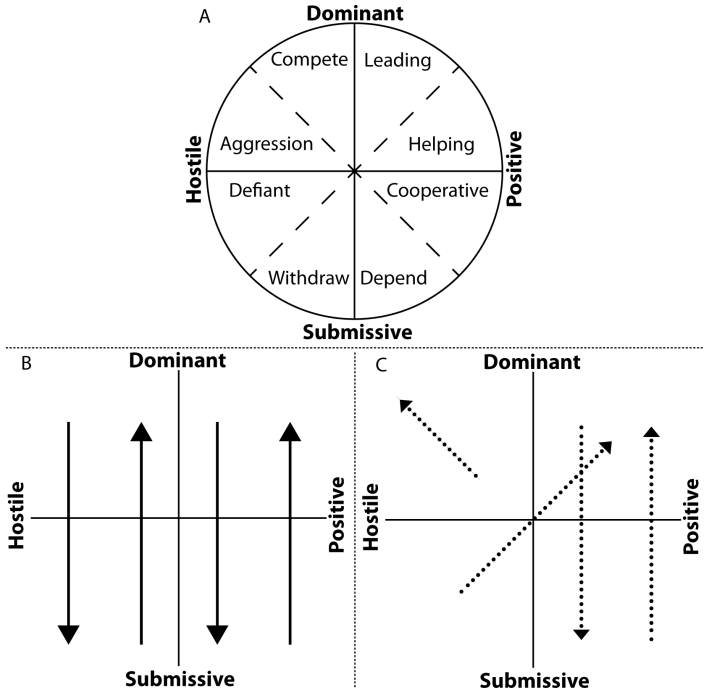

Figure 1.1: A) Leary's Rose with the different stance segments (affiliation is the horizontal axis, dominance is the vertical axis). We distinguish the segments Dominant-Positive (DP), Submissive-Positive (SP), Submissive-Hostile (SH) and Dominant-Hostile (DH). B) The relation between behaviour of the interactors. Dominant behaviour is met with submissive behaviour and vice versa. Positive behaviour is met with positive behaviour and hostile behaviour is met with hostile behaviour. C) The alternative relation between behaviour of the interactors as proposed by Ordford [144].

1.3.2 Role-Play in Training

Currently, the Dutch Police Academy offers training sessions where actors play the role of a suspect. The actor plays a suspect from a scenario that is based on a case. The suspect is described as a persona in the scenario. After the role-play, the actor

describes what effects the actions of the student had on the suspect. The actor uses terms from psychological, social, and interpersonal theories.

1.3.3 Role-Play in Future Training

We want to use a virtual human that plays the suspect in a police interrogation. Virtual humans in social skill learning offer learning by experience; the student can experience a social interaction. Using virtual humans to train students in social skills is not a new idea. There are many examples of virtual humans used in social skill training and some in the interrogation domain. For example in [163, 196] a virtual Arabic civilian is questioned by US military personnel to hone their interrogation skills. The user's utterance is classified on dialogue features to detect the principal dialogue move, topic, and degree of politeness. These influence the emotions of the virtual agent, features relating to respect, bonding, and fear. These influence the compliance of the virtual agent, which in turn influence the verbal response of the agent. This closes the loop of interaction between the human user and the virtual agent: the user asks something, which the virtual agent interprets and this influences how the agent responds, which is interpreted by the human and then the interaction cycle continues. Afterwards reflecting on the interaction can provide reflective learning for the trainee, particularly when this is a reflection on his or her own interaction [27]. A virtual suspect that can provide experiential and reflective learning has to be able to provide information on the interaction it had using terms the student understands.

To make a compelling virtual suspect that can do the above, we need to know how to relate the behaviour of our virtual suspect to the behaviour of the trainee in a way that is consistent with the personality of the persona the virtual suspect is playing. The virtual agent needs to model the dynamics of such interpersonal relations. Ideally, the agent can analyse the speech and non-verbal messages of the trainee to determine his or her level of friendliness or aggression and use these interpretations to update their interpersonal relationship values. The response of the virtual suspect is then based on the interpersonal status of the interrogator (e.g. if you make him angry...) and the suspect (...he will respond angrily). The virtual suspect's response has to be comparable to, or can pass for, 'human suspect behaviour' in such a way that the human interactant is capable of understanding the virtual suspect's interpersonal status otherwise the user cannot learn interpersonal skills from the interaction. In the rest of this thesis, I shall describe our efforts towards these goals.

2

Social Skill Training Systems

The idea of using interactive systems to train social skills is by no means new. In this chapter I will describe some of the systems that could be viewed as predecessors to our Virtual Suspect.

2.1 Introduction

The work in this thesis is interdisciplinary as working towards a virtual agent for social skills training touches upon several research fields, for example, affective computing, intelligent virtual agents, and serious gaming. Yet social science and psychology also play an important role in this thesis, in particular research on police interviewing and the development of interpersonal relations. Throughout this thesis I refer to the literature where appropriate, but in this chapter I will try to explain how my work relates to and fits in the broader field of social skills training with interactive virtual human systems.

From a historic perspective the work in this thesis is influenced by Argyle [4] who proposed that human social behaviour is created of smaller elements of behaviour. These elements are combined to create a communicative message for other social beings to observe. The notion that such a message, consisting of concrete elements, can be constructed with the intention of conveying information in an interaction clearly links to Goffman [71] and his "Interaction Rituals". He analysed interactions as abstract entities, removed from the context, place, person, and time. His conclusions were that humans employ their social behaviour as normative judgements of everything and everyone in the interaction and that once a judgement is taken up it cannot be easily revoked. Revoking a judgement that a person has previously committed to, would be considered a loss of face (an embarrassment) not only for himself but also for those in the interaction that have thus far agreed with the judgement. When his judgement is called into question, for whatever reason, a person has to consider which face to maintain. Does he threaten his and his inter-

actant's face by changing the judgement or not? Goffman's realisation was that most people will realise that such a face-threat is present and will consciously construct a social situation where the threat is averted. The implication of the work by Argyle and Goffman is that an interaction is something that can be systematically investigated, abstractly described, reasoned about, and deliberately constructed. A realisation that brought social interactions within the realm of inanimate computational machines.

2.2 Interactive Social Virtual Humans

Automatically generated social behaviour is important for human-like interaction with virtual humans [205]. Models of human behaviour can be used to make the behaviour of artificial characters more believable to humans [179]. Believable virtual humans are being employed in systems for training social skills and the application areas differ from the skill required for successfully conducting a negotiation (e.g. [186]) to the skills needed to successfully get through a job interview (e.g. [201]). While the interaction with virtual humans may elicit some degree of learning by itself through exposure, the most important part of the experience lies in the reflection on what has happened [108]. Along the same lines, *explainable artificial intelligence* [49] advocates the use of virtual humans that can explain their reasoning. These explanations of actions taken by the virtual humans could be used to improve the user's learning process, especially when the virtual humans clarify their actions in terms of the theories a user has to understand. A simple example of such a clarification might be a virtual human saying "*I started shouting because your competitive stance made me angry.*"

2.2.1 Systems for Social Skill Training

A number of research projects aim at the development of virtual characters for social skills training and serious games. For quite some years the Virtual Humans project at the Institute for Creative Technologies (ICT) has been building embodied agents that are integrated into training environments for learning interpersonal skills [103]. The mission rehearsal exercise (MRE) [87] is a training system in use by the military and uses virtual characters in a virtual world to allow trainees to practice high stress interactions and situations only available through simulation. Another example is teaching negotiation skills using artificial conversational agents in a system similar to MRE [197]. Also, the people at ICT have used virtual humans for training negotiations between a commander and a citizen in a law enforcement setting [194]. These training programs are set in a military context, so their scenarios are not suitable for civilian or police trainees. Additionally, they do not use interpersonal stance theory, which is explicitly used in police training. However, many of the systems they have developed are suitable for our goal, for example the Virtual Human Toolkit [83]. "The ICT Virtual Human Toolkit is a collection of modules, tools, and libraries designed to aid and support researchers and developers with the creation of virtual human conversational characters."[1] It is worth mentioning that there are other initiatives that work towards virtual agent systems that are easy to use for researchers and

[1] Retrieved from vhtoolkit.ict.usc.edu, May 2016.

developers. One example is the Articulated Social Agents Platform (ASAP) [110] that focusses on real-time incremental behaviour generation. Another example of such a real-time system is GRETA [136, 149]. The SEMAINE project[2] built a Sensitive Artificial Listener (SAL), which is a multimodal dialogue system with the social interaction skills needed for a sustained conversation with a human user [19], and the SAL uses modules from the GRETA agent.

The work in this thesis is not unique in its focus on police interviews. As early as 1997, Olsen [139] described a system for training police interviewing by means of interaction with a computer simulated suspect. Later also, Luciew et al. [124] built virtual learning systems to train police officers in interviewing children who have been victims of sexual abuse and to train police officers to interrogate suspects on that matter. In fact very recently, in a pilot study Bosse and Gerritsen [26], used a virtual character in a serious game that aimed to train students of the Dutch police academy in aggression de-escalation skills! They report that "participants from a possible group of end users sees the potential of simulation as an instrument for communication training." The end users suggested improvements on presence by, for example, increasing the emotional impact and intimidation by the virtual character, and creating better conversations as they were too 'aware they were only playing a game'.

In other domains than the safety and security domain, virtual agents are also used to train social skills. Work that uses interaction stance theory is the Delearyous project [201]. They developed an interactive conversational agent system for practising professional conversational skills based on Leary's interaction stance. Based on transcriptions of conversations between a human and a virtual character playing the role of an employee in a negotiation between a manager and an employee, their system was able to automatically annotate text on stance and respond appropriately [199, 200]. However, the Delearyous training system is set in a civilian (office) context and not in police training, which means that their recognition might not be sensitive to the specific dynamics of police interviews. Further, they focused only on stance recognition and not on stance computation and expression. The aim of the EU TARDIS project[3] was to develop an embodied conversational agent that acts as a virtual recruiter to train youngsters to improve their social skills [3, 97]. In that context, Chollet et al. [46] considered the analysis of sequences of non-verbal behaviours and expressions of interpersonal attitudes. They generated new sequences of non-verbal behaviour expressing interpersonal attitudes, by using a sequence mining technique on an annotated corpus filled with such behaviours. Callejas et al. [39], in that project, created a computational model for social attitudes for the virtual recruiter for job interview training. The "FearNot!" storytelling application can create an interactive virtual drama game with the topic of bullying in a school [7, 8]. Campos et al. [40], in the SIREN project[4], developed a virtual agent that can engage in natural conflict situations. This agent was embedded in a serious game intended to help children learn conflict resolution strategies. Cordar et al. [48] used social virtual

[2] www.semaine-project.eu
[3] www.tardis-project.eu
[4] www.sirenproject.eu

agents to train conflict resolution skills with healthcare professionals. They found that "virtual humans can be used to teach real humans best communication practices that humans can potentially apply in real-world situations". Morina et al. [132] showed that virtual reality technology that incorporates social interactions may be successfully applied for therapeutically treating social anxiety disorder by exposure to feared real-life social situations. Hartanto et al. [82] showed that such exposure treatments of social phobia with virtual reality and virtual agents might even be possible in the comfort of the patient's home. Op den Akker et al. [142], created a virtual coach that monitored the physical activity of the user and that presented feedback messages. Impressively, they managed to get a virtual agent working on a smart-phone using a slimmed down version of the ASAP realiser [106, 110].

Damian et al. [53] investigated whether there is a difference between the effectiveness of traditional training and training with interactive virtual agents. They compared job training with a virtual human to a traditional method with a book and coach. Pupils who trained with the virtual human rated themselves statistically significantly less nervous before a job interview, compared to pupils who trained with a traditional method. As well as that, professional career counsellors rated the pupils who trained with a virtual human as performing better at job interviews than the control group with the traditional method.

Realistic behaviour of a virtual human can elicit learning in a user by experiencing the interaction. Architectures for artificial social agents often place emphasis on the reasoning (goals, planning, actions), emotion (appraisals, mood, emotion), and dialogue (grammar, utterances) of an agent (e.g. [56, 127, 180]). All this to increase the 'positive things' in an interaction with the user: affiliation, cooperation, respect, coordination, understanding, and so on. However, for a learning application it can be beneficial to have a virtual character decrease these 'positive things' in an interaction to facilitate learning by making mistakes [167]. A virtual agent can allow the user to make mistakes by being non-cooperative. However, the virtual human needs to do more than simply refuse to give in or show behaviour that the user was tasked to prevent [194]. The virtual human should consider the goals that fit the role it is enacting and the goals of the tutoring application in which it serves [30]. In a training application it is important for the system to have the ability to explain its reasoning [49]. Such 'explainable intelligence' can lead to learning by reflecting on the interaction [108]. We try to create a model that can provide the information needed to explain its behaviour. During the interaction the model has states and state transitions, a log of these transitions provides information on the interaction that the user had. The user can use this information to evaluate his interaction as it provides insight into why the interaction went the way it went. For example, the user could compare his intentions with the way the virtual human interpreted his intentions.

2.3 Models that Compute Responses

A *response model* is a model that computes which response a virtual character should give. A model that is based on how humans behave can be used to make the behaviour of a virtual agent more believable to humans [179]. The state of the response model

(RM) can be used to select the most appropriate behaviour in the virtual human's repertoire (e.g. make a sad face and say "*You're not nice!*"). Then the human can respond to the virtual human and the cycle continues. A response model should take into account the specific role that the character plays. In this thesis that is a suspect with all the tactics and psychological manoeuvring that are involved. Whenever a virtual character says something, it has to make a choice how to say it. This 'how to say it' is reflected in many aspects of behaviour, for instance the choice of words [172], the turn-taking [153], and the facial expression [151]. Walker et al. [206] called this Linguistic Style Improvisation. An RM works independently from components that compute the content of the message of a virtual agent, which is the 'what to say'.

Reisenzein et al. [159] discuss how computational modelling of emotion benefits from the exchange of ideas and practices from psychology and computer science. They propose that emotion theories should be deconstructed into their basic assumptions to be able to construct a more unified or standardized conceptual system or implementation. We are interested in the interpersonal and social workings of an interaction (in a police interview) and do not focus on emotion. However, the idea of deconstructing social and interpersonal theories into their basic assumptions has beneficial results. In chapter 5, we describe what we include in our response model for the virtual suspect agent and how that is based on observed interactions in police interviews. The interpersonal concepts we include were selected by deconstructing the social theories that describe a police interview into the basic concepts from these theories.

According to Ochs et al. [138], before their work, computational emotion models of social character in social context were mostly static and did not consider the dynamics of the social context. They go on to propose a model, aimed at the improvement of NPC credibility in video games, of the *dynamics* of a character's emotions based on the social relations and the personality of the character. Based on the literature (among others [185]) Ochs et al. [138] consider four social variables to represent a social relation: degree of liking, dominance, solidarity, and familiarity. The literature that such models are based on often deals with 'standard' people in 'standard' situations. Whatever standard might mean in this case, suspects and police interrogations are not standard and it is questionable whether all theories based on 'standard' people are appropriate in all special circumstances.

Taylor [188] investigated communication behaviours in crisis (hostage) negotiations in order to formulate a model of the interrelationships of the communicative behaviours of suspects and negotiators. A hostage negotiation is not entirely unlike a police interrogation, mainly they can both be very stressful situations and for the suspect it is often an abnormal situation to be in. Taylor found a "cylindrical structure of communication behaviour, revealing 3 dominant levels of suspect-negotiator interaction (Avoidance, Distributive, Integrative). At each level of the structure, interactions were found to modulate around 3 thematic styles of communication (Identity, Instrumental, Relational), which reflected the underlying motivational emphasis of individuals' dialogue" [188, p4].

Characters that are not compliant in interaction, such as suspects, require a different computational model of the mind than is often pursued in more classical ar-

tificial intelligence. A non-compliant character does not try to "reach some form of optimal behaviour as measured by an agent/centred objective performance criterion, but really to produce behaviours which are credible from the player's standpoint" [138, p281]. Roque and Traum [163, 194] distinguish three levels of compliancy: compliant, reticent and adversarial. "When characters are compliant, they provide information when asked, but fall short of Gricean cooperativity because they do not provide helpful information that was implicated rather than explicitly solicited. When characters are reticent, they provide neutral information, but will evade any questions about important or sensitive information. When characters are adversarial, they provide deceptive or untruthful answers" [194, p67].

Olsen [139] describes a system that can teach police students to build rapport while maintaining professionalism, listen to verbal cues and detect important changes in both verbal and non-verbal behaviour. A list of 400 predefined questions are available for the police officer to chose from. The simulated suspect responses are given based on the question and the internal state of the suspect. The internal state consists of the mood of the suspect (angry, denial, or compliance) and the rapport between the suspect and user. Olsen used a vector to describe the mood of the suspect in these three states: (P_{anger}, P_{denial}, $P_{compliance}$). For example, a question that was appropriate would have a low P_{anger} and P_{denial} but a high $P_{compliance}$ vector attached to it, thus influencing the mood of the suspect in a positive way. Luciew et al. [124] built an interview and interrogation immersive learning simulation, specifically to train police officers in interviewing children who were victims of sexual abuse and interrogating suspects on that matter (i.e. two prototype systems were developed). In this system the behaviour of the agent is dependent largely on the proficiency of the user in detecting non-verbal cues and reporting them outside the interaction. The topic of the questions seems to be the only direct influence the user has on the behaviour of the agent during the interaction.

In op den Akker et al. [141], we reported on the results of annotating stance of human role playing suspects and police officers in a training setting[5]. We found that annotating stance is a hard task and that it is difficult to get satisfying inter-rater agreement when rating on an utterance level, yet that it is possible to get predictable patterns using a majority vote annotation. Behaviours have different meanings depending on the place in a temporal sequence of behaviours and the role of the speaker. Novielli and Gentile [134] investigate the recognition of the interpersonal stance of the user when having a dialogue with an ECA used as an interface agent. Chollet et al. [45] show recognition of the stance taken by the interviewer so that the virtual human can respond to it.

[5] See also chapter 3.

Part II

Suspect Behaviour Analyses

Part II: Suspect Behaviour Analyses

In this part of the thesis, I will present our work on analysing the behaviour of suspects in police interrogations. We collected a corpus of practice police interrogations: the Dutch Police Interview Training (DPIT) corpus. In this corpus professional training actors play the role of suspects who are being interrogated by students of the Politieacademie (Dutch Police Academy). We used this corpus to inform the behaviour of a virtual character that will play the role of suspect in a practice police interrogation. Students of the police academy will use this 'virtual suspect' to practise their social skills. First we need to know whether observers can agree on what interpersonal stance is taken in the corpus. In chapter 3 we will state that stance is a fuzzy notion but we found that a majority vote of multiple observers can indeed reveal the dynamics of stance taking. The question remained whether the proficiency of the actor has an effect on the reliability with which observers interpret the actor's actions. In chapter 4 we will show that observers agree better on some stances, that some actors are better at portraying an interpersonal stance, and that validity and agreement do not always go hand in hand. Based on these results we are confident we can use a corpus of acted interrogations to inform the creation of a cognitive response model for the virtual suspect. In chapter 5 we lay the groundwork for such a model by creating a 'semantic model' of the interactions observed in the corpus.

3

Interpersonal Stance in Police Interviews: Content Analysis

Building artificial conversational characters that play the role of a suspect in a police interrogation game requires computational models of police interviews as well as of the internal psychological mechanisms that determine the behaviour of suspects in this special type of conversations. Leary's interpersonal circumplex is used in police interview training as a theoretical framework to understand how suspects take stance during an interview and how this is related to the stance and the strategy that the interviewer takes. Interpersonal stance is a fuzzy notion. The question that we will consider in this chapter is whether different observers of police interviews agree on the type of stance that suspect and policemen take and express in a face-to-face interview. We analysed police interviews and we will report about a stance annotation exercise. We concluded that although inter-annotator agreement on stance labelling on the level of speech segments is low, a majority voting 'meta-annotator' is able to reveal the important dynamics in stance taking in a police interview. Then we explored the relation between the stance taken by the suspect and turn-taking behaviour, overlaps, interruptions, pauses and silences. Our findings contribute to building computational models of non-player characters that allow more natural turn-taking behaviour in serious games instead of the one-at-a-time regime in interview training games[a].

[a]This chapter is based on: H.J.A. op den Akker, M. Bruijnes, R.M. Peters & T. Krikke. *Interpersonal stance in police interviews: content analysis.* Computational Linguistics in the Netherlands Journal, 3:193-216, 2013

3.1 Introduction

In a serious game a police trainee can interrogate a conversational character that plays the role of a suspect. Building such a tool requires valid analysis of police interviews. Interpersonal stance is a key construct that is used in training suspect interviews. In the Netherlands at the Police Academy Leary's two-dimensional circumplex [116], also known as Leary's Rose, is used as a framework to describe and understand how the interlocutors respond to each others' stances [202]. The relation between stances taken in a police interview not only shows in the words being spoken, in postures and facial expressions, but also in the timing of speech, in interruptions and silences [140]. We are interested in building computational suspect models that underlie behaviour (e.g. turn-taking), so that the artificial suspect shows believable and natural (turn-taking) behaviour that fits the stances taken in the course of the interview [98, 191]. In order to analyse the relation between suspects (turn-taking) behaviours, and stance taking we transcribed speech and annotated stance of interlocutors in a number of video recorded police interviews collected at the Dutch Police Academy.

In this chapter we will report about this annotation work where a number of annotators labelled the interlocutors' turns with categorical stance labels. We will discuss the results in terms of inter-rater agreement. What are we annotating when we annotate stance in police interviews? In Section 3.2 we will discuss the model that we use as basis for our stance annotation scheme and in Section 3.3 we will review related work in stance annotation as well as in discourse analytical studies of various styles of police interviews that are employed. Since stance is a fuzzy notion we can expect that annotators will disagree quite often if we force them to make a single choice from a fixed set of stance labels. In Section 3.4 we will present our annotation results. We will argue that, despite a low inter-rater agreement, using a majority voting system with a number of annotators we are able to identify the agreed global dynamics in stance taking over the course of an interview. Fuzziness of labels is one of the causes of a low inter-rater agreement. In Section 3.5 we will present a computer simulation to get an idea how the fuzziness of the stance labels used in the annotation influences our reliability measure. Then, in Section 3.6 we will discuss some fragments from our corpus focusing on turn-taking and we will formulate and present hypotheses about the relation between strategies, stance and turn-taking behaviour in police interviews. In Section 3.7 we will conclude with a reflection on our findings and we will formulate some challenges ahead.

3.2 Interpersonal Stance

According to the English dictionary[1], stance is either posture or attitude, "the way in which someone stands especially when deliberately adopted". Stance and stance taking are subjects of research in social psychology, in social linguistics and, more recently, in technology oriented social signal processing circles. "One of the most important things we do with words is take a stance." [60, p.139]. For DuBois stance is realized usually by a linguistic, that is to say, a social act. People take stance inter-

[1] Oxford English Dictionary

actively. Stance has three aspects: evaluation, positioning and alignment. The stance taker *evaluates* (assesses or appraises) something or someone ("that's horrible"), he *positions* himself towards something, a situation or someone ("I don't like that"), in alignment or dis-alignment with others (agreement, disagreement). Stance can be affective or epistemic or both. DuBois proposes a model of stance with three components: the stance taker, the object of stance taking, and the stance that the stance taker is responding to. DuBois as well as Karkkainen considers inter-subjectivity an essential ingredient of stance taking. "Stance is not primarily situated within the minds of individual speakers, but rather emerges from dialogic interaction between interlocutors in particular dialogic and sequential contexts." [99, p.700]. The notion of stance is explored by Chindamo et al. [44] through a review and discussion of some of the relevant literature. In line with the studies mentioned above the conclusion drawn by Chindamo et al. is that: "studies of stance and stance taking should (therefore) focus both on the expression of a speaker's stance and the reaction it leads to in his/her interlocutors." Scherer analyses interpersonal stance as a particular affect category and provides the following characterization [170, p.705-706]: "The specificity of this category (of stance) is that it is characteristic of an affective style that spontaneously develops or is strategically employed in the interaction with a person or a group of persons, colouring the interpersonal exchange in that situation (e.g. being polite, distant, cold, warm, supportive, contemptuous). Interpersonal stances are often triggered by events, such as encountering a certain person, but they are less shaped by spontaneous appraisal than by affect dispositions, interpersonal attitudes, and, most importantly, strategic intention."

We are particularly interested in the interpersonal stances that the interlocutors deliberately or 'automatically' take in the encounter of a police interview. Holmberg [88] analyses the function of stance in police interviews. He actually uses the term 'attitude': "the psychological tendency to evaluate and express a positive or a negative value with regard to a certain attitude object" [88, p.37]. Negative attitude generates avoidance, positive attitude serves an approaching function. Many police officers have been exposed to stressful events that may cause a negative attitude towards serious crime suspects, causing interview practices that are characterized by dominance and hostility. In training conversational skills and strategies police trainees in the Netherlands use Leary's theory of interpersonal relations as a framework for analysing their own behaviour (see chapter 1). They learn to understand the suspect's behaviour as a response to their own behaviour as an expression of stance taking. Leary's model is known as the interpersonal circumplex or under the more popular name *Leary's Rose* [116]. It is presented by a circular ordering of eight categories of interpersonal behaviour, which is situated in a two-dimensional space that is spanned by two orthogonal axes, representing the two "basic dimensions of interpersonal behaviour" [104, p.5]: affiliation (friendliness versus hostility), the horizontal axis, and power (dominance versus submission), the vertical axis. Accordingly, every form of interpersonal behaviour is determined by the amount of affiliation and by the amount of dominance towards the other. Leary [116] formulated the principle of 'reciprocal interpersonal relations': "any interactional act is designed to elicit from a respondent reactions that confirm, reinforce, or validate the actor's self-presentation and that make it more

likely that the actor will continue to emit similar interpersonal acts." [104, p.6]. Two conversational partners are influencing each other with their stance during a dialogue ('interpersonal reflexes').

Psychological research with the interpersonal circumplex model has demonstrated the value of that model for integrating a broad range of psychological topics. Rouckhout and Schacht [164] present the results of a Dutch study with the purpose (1) to find out whether there is a circumplex structure underlying a comprehensive set of Dutch interpersonal adjectives, and (2) to construct a set of Dutch interpersonal circumplex scales. They found a set of Dutch adjectives for each of the eight categories of Leary's Rose. These frequently recur when people describe the interpersonal stance of actors in a personal encounter. Table 3.1 (top) shows the Dutch adjectives scale from Rouckhout and Schacht [164]. Table 3.1 (bottom) shows a similar set of English adjective scales from Wiggins [212]. Understanding of interpersonal behaviour requires study of both the linguistic and the non-verbal levels of human communication. During the annotating of police interviews we use both scales for deciding which labels best fit the observed stances (see Figure 3.1).

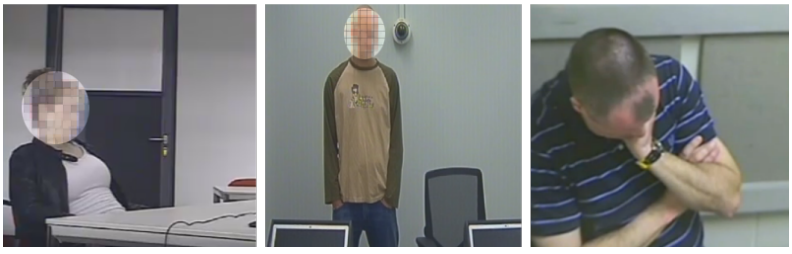

Figure 3.1: Stances taken by suspects during a police interview. (Faces blurred for privacy).

3.3 Related Work

3.3.1 Interpersonal Stance Annotation

Compiling a reliable corpus of emotion annotated dialogues is hard. There are two difficulties that we aim to address in this thesis: (1) Gathering such a corpus often involves using actors but acted emotions might not be the same as spontaneous emotions. It can be difficult to gather a corpus of spontaneous emotions and this is especially true for negative emotions [13]. The question remains are acted emotions a viable substitute for 'real' emotions? In chapter 4 we will investigate this by looking at the effect actor proficiency has on the recognisability of the portrayed emotional behaviour. (2) Many emotion researchers have discussed problems involved in annotation, such as the choice between a categorical forced choice annotation scheme versus a one or multi-dimensional continuous scheme, for example, Craggs and McGee Wood [50], Busso and Narayanan [37]. Here, one of the problems is the low inter-rater agreement due to the subjectivity of the perceptions of emotions. Similar issues arise if we want to study and annotate interpersonal stance in dialogues. The 'emotion

Table 3.1: Top: Dutch adjectives scales for the categories of the interpersonal circumplex from Rouckhout and Schacht [164]. Bottom: English adjectives scales for the categories of the interpersonal circumplex from Wiggins [212].

Compete	**Aggression**	**Defiant**	**Withdrawn**	**Depend**	**Cooperative**	**Helping**	**Leading**
Eigenwijs	Onbarmhartig	Afhankelijk	Verlegen	Pretentieloos	Minzaam	Ongedwongen	Onwankelbaar
Cynisch	Geniepig	Onbeholpen	Onderdanig	Eenvoudig	Lief	Charmant	Extravert
Gehaaid	Pretentieus	Gezagsloos	Schuchter	Moederlijk	Tolerant	Mededeelzaam	Praatgraag
Aanvallend	Wreedaardig	Tactloos	Stil	Braaf	Liefdevol	Vrolijk	Krachtig
Vrijpostig	Grof	Onpersoonlijk	Beschaamd	Doodernstig	Inschikkelijk	Spontaan	Geestdriftig
Autoritair	Bevooroordeeld	Kortaf	Gereserveerd	Nederig	Zachtmoedig	Opgewekt	Geraffineerd
Dominant	Doortrapt	Asociaal	Twijfelend	Onschuldig	Barmhartig	Galant	Hardnekkig
Vechtlustig	Brutaal	Egocentrisch	Timide	Overbezorgd	Bedaard	Attent	Fier
Impulsief	Achterdochtig	Liefdeloos	Introvert	Buigzaam	Gewillig	Voorkomend	Resoluut
Onstuimig	Tegendraads	Onoprecht	Terughoudend	Weekhartig	Onbevooroordeeld	Sociaal	Ambitieus
Ongegeneerd	Sluw	Bedrieglijk	Gesloten	Alledaags	Bescheiden	Vriendelijk	Spraakzaam
Dikhuidig	Geslepen	Oneerbiedig	Schuw	Onopvallend	Teerhartig	Loyaal	Kordaat
Onbedeesd	Listig	Arrogant	Gemaakt	Bedeesd	Zachtzinnig	Tactvol	Vol vuur
Onverlegen	Despotisch	Afgunstig	Afzijdig	Volgzaam	Plooibaar	Menslievend	Vastbesloten
Onbeschroomd	Schaamteloos	Intolerant	Naïef	Meegaand	Teergevoelig	Behulpzaam	Levendig
Firm	Crafty	Coldhearted	Unsparkling	Forceless	Unwily	Charitable	Perky
Dominant	Cunning	Cruel	Introverted	Unauthoritative	Uncunning	Tender	Enthusiastic
Forceful	Boastful	Unsympathetic	Timid	Unbold	Unsly	Sympathetic	Outgoing
Domineering	Wily	Warmthless	Bashful	Unaggressive	Softhearted	Kind	Extraverted
Cocky	Calculating	Uncheery	Shy	Unargumentative	Accommodating	Cheerful	Self-assured
	Tricky	Unneighbourly	Meek	Undemanding	Gentlehearted	Friendly	Self-confident
	Sly	Distant		Uncalculating	Tenderhearted	Neighbourly	Assertive
	Ruthless	Dissocial		Uncrafty		Jovial	Persistent
	Ironheaded	Unsociable		Boastless			
	Hardhearted	Antisocial					
	Uncharitable						

classification task' based on Leary's Rose was introduced in Vaassen et al. [200] and executed in the Belgian project deLearyous. Annotation work can have two different goals: (1) content analysis, aiming at finding correlation between various aspects of the content, for example to study dependencies between stance taken by the police interviewer and the response stance taken by the suspect or (2) to build a train and a test corpus for machine classification. The aim of Vaassen et al. was the latter. They report about the performance of a number of machine classifiers for the task of classifying the stance expressed in the words spoken by the human interlocutor when interacting with the virtual non-player character in a serious game. Our aim is of the first type and related to building a computational model of the virtual suspect and his verbal and non-verbal behaviours when being interviewed.

In the deLearyous project [201] the focus was on the machine classification of stance in written dialogue. To investigate the quality of their stance annotations, four annotators labelled a small subset of sentences from the corpus. The inter-annotator agreement was calculated using Fleiss' kappa and was found to be $\kappa = 0.29$ over the eight stances, and $\kappa = 0.37$ on a quarterly metric ('Leading equals Helping', etc.)[200]. Their low kappa scores are similar to the kappa scores we found in the current study and a further indication that identifying the position of a speaker on the interpersonal circumplex is a difficult task. However, they state "since the goal of the application is to simulate human behaviour, these results also imply that it is not critical for the final application to reach a perfect level of prediction. In fact, due to the subjective nature of the annotation process, an objectively 'correct' result likely does not exist." [201, p.5]. Using machine learning techniques to perform the stance classification task, Vaassen et al. [199] managed to achieve an accuracy of 52.5% on the classification of the stance quadrants. This score means that their classifier can correctly label one out of two sentences into the correct quadrant in Leary's Rose, a result that might not be sufficient for a social communication training tool. However, by using more context information the classifier's accuracy could be increased sufficiently to have a convincing artificial conversational agent [209].

Burkett et al. [36] describe the results of stance annotation of textual chat interactions in an educational game setting using Leary's Rose. Their goal was to see if personality traits can be detected automatically from dialogues and what personality traits are most prevalent over the course of the game. Six categories of Leary's Rose were used for the coding scheme: the Helping and Co-operative categories and also the Aggressive and Defiant were categorized into one. Statements that did not fit into any of the categories were coded as neutral, indicating that there was no evidence of any of the six categories present. Two independent raters annotated a corpus of 1,000 excerpts with an average kappa of 0.65.

Allwood et al. [2] analysed stance taking and its relation with conflict in political television debates. They define stance as "an attitude which for some time is expressed and sustained in communication, in a unimodal or multimodal manner." Where attitude is taken as "a complex cognitive, emotive and conative orientation towards something or somebody" [2, p.1]. Their qualitative analysis of a number of conflict episodes showed that some clusters of stances co-occur. Three stances were found to be characteristic for conflict episodes: aggressive, provocative, resig-

nation. The latter distinguishes from the first two in a number of expressive features, quiet voice and non-focused gaze. Other behaviours such as overlap, interruption and raised voice are less unique for types of conflict related stances.

3.3.2 Analyses of police interviews

Police interviews are analysed by psychologists and sociolinguists interested in the effectiveness and characteristics of various interview styles and interrogation tactics. Special attention is paid to the interaction between interview style and the suspect's willingness to talk freely, to admit or to deny.

Benneworth [16] performed a discourse analytical study of UK police interviews of suspected paedophiles. Interruptions by the interviewer are a prominent phenomenon of the commonly used interrogative and accusatory interview style. Benneworth highlights the importance of encouraging uninterrupted narratives from suspects.

Jones [96] studied differences between Afro-Caribbean and White British suspect interviews in the UK. She focussed on overlapping talk and found differences in the uptake of the interrupted talk; the Afro-Caribbean suspects propositions were taken up to a lesser degree than those of any other group. This clearly shows, according to Jones, that "the police officers had more power and control than the Afro-Caribbean suspects in these interviews and [that] potentially has something to do with race and suspect status".

Holmberg et al. [89] report about an explorative study among 83 criminals into the relationship between police interviewers' behaviour and suspects' inclination to admit or deny crimes. From the perpetrators' point of view they found two basic interview styles: one characterized by dominance, the other by humanity. In response to these styles the suspect will experience being respected or worried. Dominance is related to the perceptions of interviewers as aggressive, brusque, and impatient. It also relates to hostility, dissociation, and nonchalance. Humanity showed a positive correlation to feelings of respect, and a negative correlation with feelings of being condemned and anxiety [89, p.39].

Snook et al. [177] examined questioning practices of Canadian police officers. Transcripts of police interviews with suspects and accused persons were coded for the type of questions asked, the length of interviewees responses to each question, the proportion of words spoken by interviewer and interviewee, and whether or not a free narrative was requested. Results showed that, on average, less than 1% of the questions asked in an interview were open-ended, and that closed *yes or no* questions and probing questions composed approximately 40% and 30% of the questions asked, respectively. Free narratives were requested in approximately 14% of the interviews. The limited knowledge about the current, in 2012, questioning practices being utilized in interrogation rooms in North America provided the impetus for their study.

Beune et al. [17] analysed sequences of dialogue acts in police interviews with suspects from different cultures. The police acts were coded using the 'Table of Ten' strategies, a list of ten tactics for hostage negotiations proposed by Giebels [66]. Strategies are among others: 'Emotional Appeal', 'Rational Convincing', and 'Direct Pressure' (see Table 1.1). The suspects' acts were coded by three different content

categories of inform acts. The aim of the study was to see if cultural factors (in particular the difference between high and low context communicators) mediate the effect of the interview strategy and the responsiveness of the suspect in terms of the willingness to provide information about his own involvement, or about others involved in the case at hand [18, 189].

All in all we can conclude that a variety of interview strategies have been identified and described in the literature; from different perspectives and with different aims. The focus is mostly on the interviewer: the police officer. The suspect's behaviours and stance are seen as dependent on the interviewer's strategy and stance towards the suspect. The focal issue is the relation between the strategy the police officer follows (whether deliberately chosen or not) and the suspect's denial or admission. Although none of the studies we have seen explicitly use Leary's interactional circumplex to describe the stances taken by the interviewer it will be clear from the above that dominance and hostility recur as important factors in studies that describe the characteristics of the predominant interview styles. Interview styles followed are related to stance taken towards the suspect and his (criminal) acts. Styles differ in the types of questions used by the police as well as in interrupting behaviour. Police officers are trained in applying various strategies and in monitoring the influence that their behaviour has on the suspect. We expect that Leary's Rose is a valid framework to describe what is going on in terms of stances taken. But do different annotators see the same things happen regarding the stances taken by the interlocutors?

3.4 Annotating Stance in Police Interviews

We performed an annotation task in which annotators independently annotated police interviews with labels for the stance categories of Leary's Rose. The question is whether different annotators see the same stances taken by the interlocutors. In this section we explain the corpus, the annotation effort, and we present statistics on the inter-rater agreement. We argue that a majority voting 'meta-classifier' is able to give a reliable picture of the essential changes in stance over the course of a police interview.

3.4.1 Annotation Material and Task

We present the *Dutch Police Interview Training Corpus* (DPIT corpus). The DPIT corpus consists of police interviews conducted by trainees of the Dutch National Police, recorded in 2012 and 2013 at the police academy. The police officers in the corpus are novice to moderately proficient police interviewers. The suspects they interview in this corpus are professional training-actors. Due to privacy requirements, the video and audio data of the corpus cannot be made publicly available.

The corpus consists of 32 interviews from 6 scenarios (cases) with a total length of approximately 13 hours. The interviews vary in length from about 9 minutes to almost an hour. Some scenarios were enacted several times (with different students and sometimes also with different actors), while other scenarios were cut into separate interviews. In the latter, the suspect was interviewed multiple times, for example

to give the police officers time to check facts. In these scenarios the same actor is interviewed by different police officers. The scenarios are presented as follows:

Bruintjes Ms Bruintjes is suspected of having bought a stolen smartphone from her cousin. She comes across as being not very bright but knows the phone was stolen.

Huls Mr Huls is suspected of the theft of a small amount of cash from a petrol station. He is a professional training actor for the police, has financial problems and has difficulty feeding his family.

Motor Actors from this scenario performed (with criminal exemption) an actual theft of an outboard motor and they play themselves. They try to appear innocent but are instructed to admit when the evidence gets too strong to reasonably deny the crime. The police students are not aware they are guilty and treat them as any other suspect.

Remerink Ms Remerink is suspected to have stolen money from her (ex) husband's bank account. She is a full-time mother and gave up her career for their kids. Her (ex) husband is wealthy and he left Ms Remerink for another woman.

van Bron Mr van Bron is suspected of arson with the intent to kill his neighbour. Van Bron has an anti-social or bipolar disorder and has a criminal record. His girlfriend made a statement implicating Mr van Bron.

Wassink Mrs Wassink is suspected to have physically attacked her neighbour over an argument about the neighbour's dog. Wassink is a working-class mother whose world is as big as the neighbourhood she lives in and she is suspicious of people not originating from her neighbourhood (like her neighbour).

The actors are allowed to change the scenarios according to their preferences and to fill in the details as they see fit. This means that instances of interviews from the same scenario may be different, yet the police officers are always training with the same *persona*.

For our annotation task we selected an interview from the Wassink case (more about this case in Section 3.6). We pre-segmented speech into speaker turns and transcribed these. Annotators, students that were introduced into Leary's Rose and into the annotation task, independently labelled speaker turns with one of the 8 stance labels of Leary's interpersonal circumplex. Annotators used the tables of Dutch adjectives (Table 3.1, top) to help them find the best fitting stance label. If there was no clear stance expressed a *neutral* stance category was chosen with the label: Neutral. We used ELAN as annotation tool [175]. We measured inter-annotator agreement using Krippendorff's alpha, a very general method for comparing an arbitrary number of annotators allowing different distance metrics on the label set [112]. Labels next to each other on the Leary's interpersonal circumplex can be considered more similar than labels of more distant categories of the circumplex. When two annotators label one and the same speaker turn with labels A and B respectively, the distance between A and B as defined by the metrics used in the alpha statistics specifies how much we penalize for this disagreement.

3.4.2 Annotation Results

A fragment of one of the interviews from the Wassink scenario, with a length of 148 speaker turns, was labelled by 9 independent annotators. The total number of labelled items produced by the 9 annotators on the 148 speaker turns shows the label distribution:

- Neutral: 215
- Leading: 309
- Helping: 269
- Cooperative: 218
- Depend: 75
- Withdraw: 70
- Defiant: 89
- Aggression: 23
- Compete: 64

Table 3.2 shows the values of the alpha statistics for the leave-one-out groups of annotators. We computed alpha with the following distance metrics (see columns 2-5 in Table 3.2):

- Boolean metric - two labels are equal (distance is 0) or not (distance is 1). For all annotators $\alpha = 0.24$;
- Quarterly metric - two labels in the same quarter of Leary's Rose are considered equal (Leading equals Helping, etc.). For all annotators: $\alpha = 0.42$;
- Quarterly metric with neutral - same as quarterly but Neutral is now considered equal to all others labels. For all annotators: $\alpha = 0.44$;
- Quarterly metric diagonal wise - two labels in the same quarter of Leary's Rose, where quarters are the adjacent octants separated by the two diagonal lines, are considered equal (Compete equals Leading, etc.). Neutral is considered equal to all other labels. For all annotators: $\alpha = 0.22$.

The table shows that the results are quite similar for all leave-one-out groups. Moreover, even using the tolerant penalty system for disagreement defined by the Quarterly metric with neutral the alpha values are rather low with a maximum of 0.44 for the whole group and of 0.51 when we leave one out.

From Table 3.2 we can draw the following conclusions. If two annotators disagree about the stance label and one chooses label A and the other label B, then it is more often the case that these two labels are in the same quarter of Leary's Rose (as for example 'Leading' and 'Helping', see column quart+N) than that they are neighbouring but not in the same quarter (as for example 'Compete' and 'Leading',

Table 3.2: Krippendorff α values with different distance metrics for the 9 leave-one-out groups of annotators. The last two columns give the Cohen κ values for the annotator and the two Majority Vote "meta-annotators".

left-out	α (bool)	α (quart)	α (quart+N)	α (diag+N)	MajVote(f)	MajVote(l)
MB	25	43	46	23	43	38
TK	27	43	43	23	25	24
DAV	23	40	42	21	60	61
MER	24	41	43	22	49	48
NIE	25	41	43	23	45	43
SJO	23	40	42	21	57	61
SOF	27	40	51	24	29	29
STE	23	41	42	20	60	59
JAN	23	40	42	21	51	55

see column diag+N). The difference between the values for the two distance metrics is significant (paired t-test, $p < 0.01$).

What leads judges to disagree about the stance label? Formally there are two types of disagreements: noisy-like and systematic. It makes sense to analyse annotations to see what type of disagreements cause the low alpha values [157]. This is not only relevant when the aim is machine classification (machine classifiers are able to learn despite noisy disagreements [158]), but also when the aim is to find correlations between different phenomena in conversations, such as between stance and turn-taking behaviour. Low inter-rater agreement may reveal problems judges have with the semantics of the stance labels. In the next section we will analyse the effect that the vagueness of the stance labels has on the alpha statistics. Differences in annotators' personal bias for one label can be shown as follows (from [157]): Compare pairs of annotators of the same data. Filter out all pairs of labels where the judges agree. Perform a correlation test on the disagreed pairs. When one of the judges has a bias towards one particular label we will find a correlation. We performed this test for two judges: SJO and STE. In particular we looked at the use of labels Neutral and Helping. In 96 unequal pairs they used 4 vs. 44 times Neutral and 41 vs. 24 times Helping. A $\chi^2(1)$ one-tailed Fisher test shows that there is a very significant difference in the use of Neutral ($p = 0.007$) between the two annotators. Further analysis of this particular case shows that the difference is mainly in the labelling of the stance taken during back-channels [217] and short feedbacks. STE labelled them Neutral where SJO chose a label that depended on whether the feedback was cooperative or opposing. Explicit instructions in the annotation procedure can easily avoid this type of disagreement.

3.4.3 Group-wise Annotation by Majority Voting

From the group of 9 annotators we construct two Majority Vote "annotators" (MVA), a 'meta-annotator' that assigns the label that has the majority vote of the group. Since more than one label can have the maximum number of votes, we construct two Majority Vote annotators or MVAs. Given a fixed label order, the MVA1 takes the first label

that has the majority vote, where MVA2 takes the last label that has the majority vote. Since the order is chosen at random this amounts to constructing two MV annotators with random choice in case of a tie. The last two columns of Table 3.2 contain the κ values for the inter-rater agreement between the annotators and the two MVAs. The maximum value obtained is $\kappa = 0.61$.

Figure 3.2 shows the annotation of an MVA, based on a group of 5 annotators, for the complete Wassink interview. The first half of this interview was annotated by all 9 annotators; the complete interview lasted about 10 minutes and 300 speaker segments. The graph shows that the stance of the suspect changed over time from Defiant/Withdrawn to Cooperative and to Defiant again at the end of the interview. This pattern confirms the intuitions the annotators expressed when reviewing and discussing the interview. The clearly visible pattern in the majority votes indicates that using them seems to be a good way to measure what is going on regarding stance and stance changes in a police interview.

We performed a computer simulation to get an idea of a) when a majority voting system is able to detect the changes in stance over the course of an interview, and b) how the inter-rater agreement between two majority voting meta-annotators is related to the within groups agreement. Suppose we have two groups of annotators, each with k members. Each of the members labels the same t items. Both groups follow the Majority Voting Protocol and assign the label (there are 8 labels) that has the maximum number of votes in the group. Then we compute the inter-rater agreement within each of the two groups as well as the inter-rater agreement between the two majority voting groups. We simulated the annotations by a Gaussian distribution around mean values (in degrees on a circle). We did this with mean values 120, 240, and 180 degrees, for the first, second, and the third 100 items, respectively and with standard deviations 30, 60, and 100 degrees, respectively. Note that 90 degrees corresponds to a whole quarter of the Rose. The higher the standard deviation the more the majority vote will fluctuate around the truth value (the mean) and the less easy it will be to detect a real change in the stance taken. Figure 3.3 shows the result for a majority voting system based on 10 simulated Gaussian annotators. It shows a clear change of stance between the first and second part, but the change is already less clear between the second and third part.

Table 3.3 shows how the α statistics computed between the two groups depend on the number of annotators k, the number of items t and the standard deviation (degrees of the circle) of the Gaussian. The table shows that even when the within group agreement is low, still for the highest values of α (last row: $\alpha(1) = 0.40$ and $\alpha(2) = 0.39$) the agreement between the two groups that follow the Majority Voting Protocol has a moderate Cohen's κ value of 0.63.

We draw the following tentative conclusions from our findings. If we annotate stance on the level of speaker segments and we force judges to choose one of a fixed number of stance labels, we find low inter-rater agreement. Nevertheless, if we take into account the fuzzy character of the meaning of stance labels and we take the most commonly assigned label, we see a moderate agreement. This agreement seems good enough to see the global stance changes over the course of an interview.

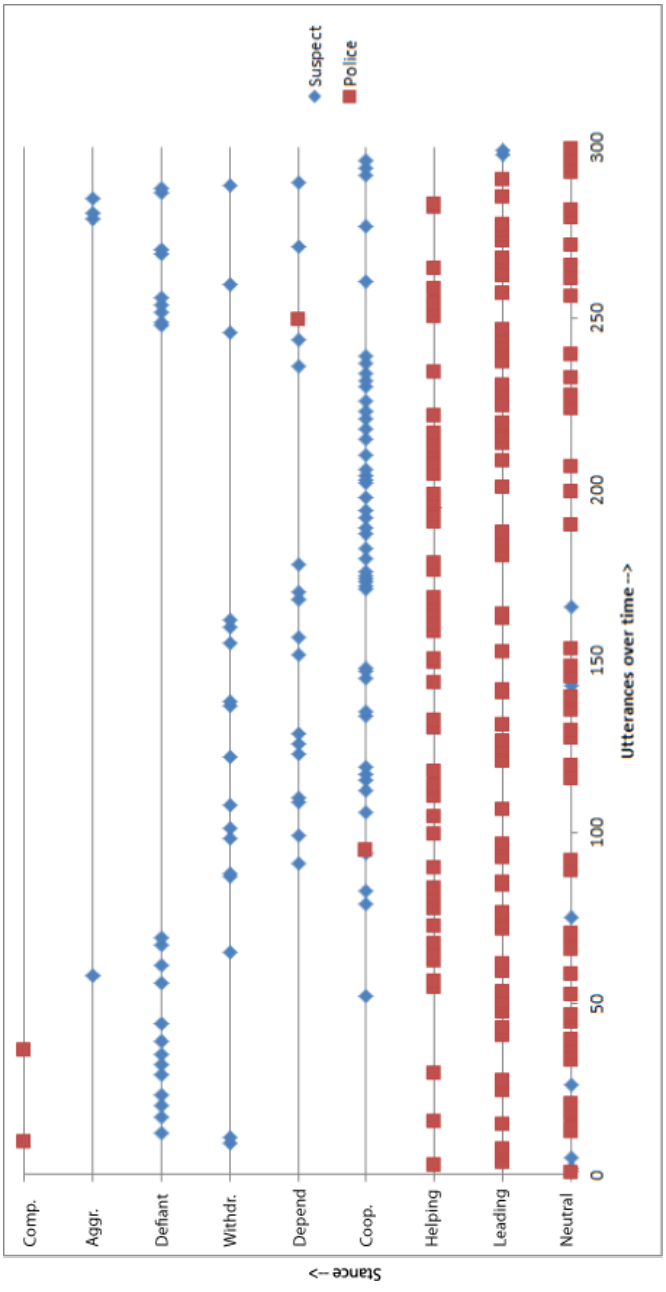

Figure 3.2: Majority voting applied to stance annotations of a police interview (group of 5 annotators). X-axis: the items/turns ordered along the time axes. Y-axes: the discrete stance label according to Leary's Rose (see Figure 1.1).

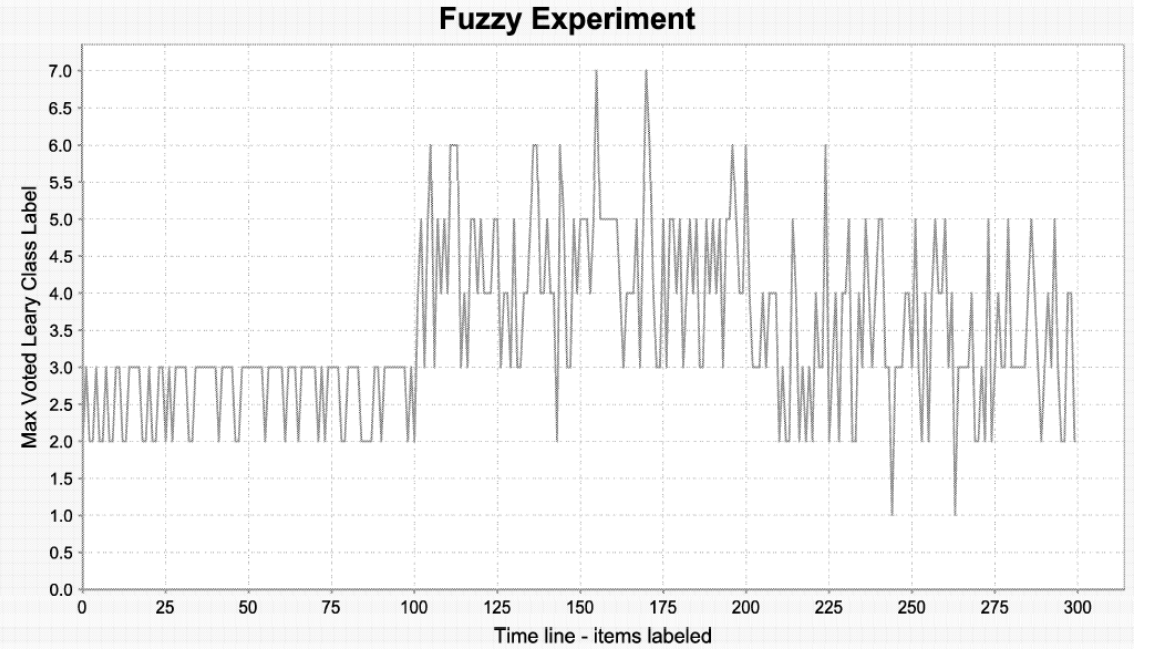

Figure 3.3: Majority vote annotation (simulated with a Gaussian distribution mean values 120,240 and 180 degrees and standard deviation 30, 60 and 100 degrees): 10 annotators simulated; 300 items; 8 labels. X-axis: the items/turns ordered along the time axes. Y-axes: the whole numbers correspond to the discrete stance label numbers according to Leary's Rose(1=Leading, etc.).

Table 3.3: The α values for two groups of annotators that labelled items following the majority voting procedure. Columns (from left to right): number of annotators per group; number of items labelled by each of the annotators and each of the groups; the standard deviation of the Gaussian distribution; the α values of each of the groups internally; the κ value between the two max voting groups. Normal (i.e. Boolean) distance metric is used.

k annos	t items	sdev	$\alpha(1)$	$\alpha(2)$	κ(betw.)
100	300	60	0.05	0.05	0.42
100	300	45	0.11	0.11	0.80
50	300	60	0.06	0.06	0.59
50	300	45	0.11	0.11	0.66
50	300	30	0.23	0.22	0.67
10	300	60	0.06	0.05	0.15
10	300	45	0.10	0.11	0.19
10	300	30	0.23	0.22	0.42
10	300	20	0.40	0.39	0.63

3.5 Simulating Annotation with Fuzzy Labels

We have seen that when we ask annotators to annotate speaker turns with one of 8 stance labels corresponding with the 8 octants of Leary's Rose, the inter-rater agreement is rather low. One of the causes of a low inter-rater agreement is the fuzziness of the stance labels. When do we call the stance that someone takes 'leading' rather than 'helping', or 'competing' rather than 'aggressive'? We forced annotators to make a choice for one of the labels. Most of the time this will be a choice between labels of adjacent octants of the circumplex. In this section we will present a computer simulation to see what the effect of the fuzziness of the adjectives is on Krippendorff's alpha measure for inter-rater agreement. In the previous section we simulated an annotator with a Gaussian distribution. Here, we are going to provide the background for this choice.

Zadek [218] modelled vague predicates by means of the mathematical notion of a fuzzy set. A fuzzy set F is defined by a membership function defined on a universe U of objects. $\mu_F(u \epsilon U)$ is a real number in $[0, 1]$, the grade of membership of u in the fuzzy set F. If u is a certain stance and F is for example 'helping' then $\mu_F(u)$ is the grade of helpingness of the stance u. In a computational model of stance u will be a sequence of feature-values; a point in a multidimensional space. Since the introduction of fuzzy sets and fuzzy logic there has been a discussion about the interpretation of this notion of fuzziness. One of the issues was the relation between the concepts of probability and fuzziness and the question whether fuzziness requires a formal logic of uncertainty that is different from the classical theory of probability. Cheeseman [43] argues that fuzziness is uncertainty about meaning and he interprets the membership function of a fuzzy set as a likelihood function. The idea comes from Loginov [123] and was the basis for constructing membership functions. Given a population of individuals (our annotators) and a fuzzy concept F each individual is

asked whether a given object u can be called F or not. The likelihood $P(F|u)$ is the proportion of individuals that answered 'yes' to the question [61].

$$\mu_F(u) = P(F|u)$$

We used this interpretation of fuzziness in our simulation experiment. We assumed that the points in the circumplex are generated according to a Gaussian distribution

$$N(\mu, \sigma)$$

with $\mu = u$, a real number in $[0, 360]$, representing a point on the circle, a certain 'objective' stance value. If this is a reasonable model we may expect that the majority vote of a sufficiently large number of annotators will coincide with the mean of the Gaussian, the 'real' stance.

The more fuzzy a concept is, the larger the standard deviation σ and the more 'confused' the (simulated) annotator is about the stance. The points generated are mapped on the vague labels 'Helping', and so on. For example when the random value u' is in $[0, 45]$, the label generated is 'Helping'. This way we generate annotations controlled by a selected stance $\mu = u$ and a chosen σ. When σ grows the inter-rater agreement will be less; there will be more confusion. In that case it will be harder to see the differences between two different stances. Consequently, it will be harder to identify changes in stance taken by people over time. Our fuzzy simulation experiment gives us some insight into how the vagueness of labels contributes to the α values for the inter-annotator agreement between the members of a group of annotators.

Figure 3.4 shows how the statistic α depends on the standard deviation. Clearly, the larger the standard deviation, the lower the inter-rater agreement. In this simulation we simulated 10 annotators each annotating 300 items with 8 different labels. The graph shows that for example when $sdev = 35$ degrees $\alpha = 0.25$. When $sdev = 45$ α becomes about 0.18.

A typical random sequence of 10 annotations generated with $sdev = 45$ and mean 157 degrees is: [Aggressive, Aggressive, Aggressive, Aggressive, Competing, Defiant, Competing, Withdrawn, Aggressive, Aggressive]. The category that has the majority vote is Aggressive, which in this case coincides with the category of the mean of the distribution. Clearly, the more annotators we have the better the majority vote will equal the mean. A final caveat is in order. The use of a normal distribution as a model for the vague meanings of the stance labels seems too simple given the outcome of our analysis in section 3.4. We found (see Table 3.2) that the distance between labels in the same quarter of the circumplex is shorter than the distance between adjacent labels not in the same quarter but in neighbouring quarters, where our model implies that they are similar.

3.6 Stance and Turn-taking

Interviews and in particular interrogations are a special type of 'talk-in-interaction' [168] in which turn-taking rules differ from those that Sacks et al. [165] formulated

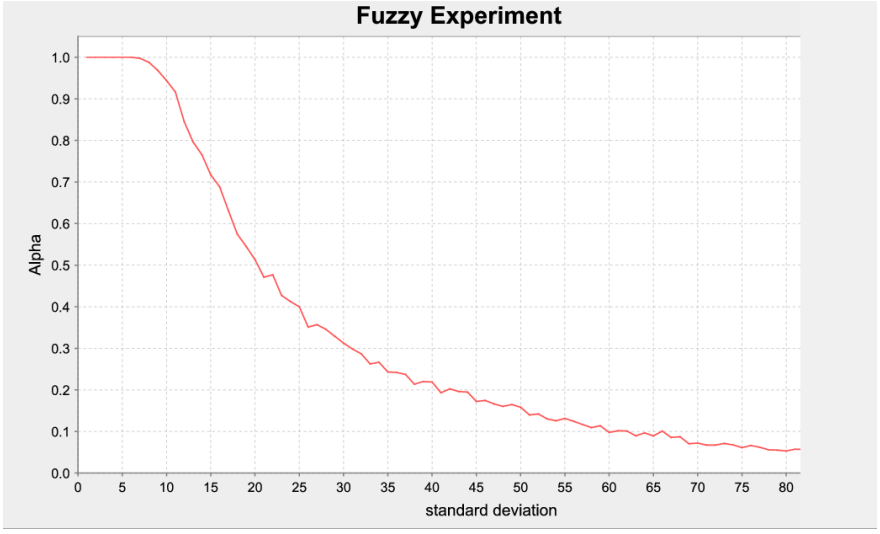

Figure 3.4: Krippendorff α values for the inter-rater agreement of 10 simulated annotators (300 items; 8 labels). X-axis: the standard deviation of the Gaussian distribution that models the fuzziness of the labels. The larger the standard deviation the more fuzzy the labels are.

for the conversation, the type of 'speech exchange system' we could see as 'normal'. In conversations turn-taking is an interactional achievement between interlocutors that are basically operating on the same level. In the emerging conversation speaker overlaps are rare and if they occur they are short (apart from short back-channels and listener feedbacks [217]). Gaps in between two speakers are also short. Moreover, exceptions are marked and need a sort of repair work. However, 'normal' conversations are quite rare. In a survey interview interlocutors have distinguished roles. Basically, the interviewer is asking the questions, the interviewee answers. Role and status determine to a great extent who gets the floor [143]. Police interviews and in particular interviews of a *suspect* are a special type of interviews and differ from survey interviews in that the interviewee is often not very willing to cooperate. Indeed, the various 'interview strategies' (empathic, investigative, dominant) that the police officer employs result in a variety of dialogue types some of which hardly deserve the name 'interview'. Each type has its own turn-taking style. In this Section we will explore how stances taken by the interlocutors are related to the turn-taking phenomena, in particular to the two observable phenomena: overlapping talk and silences. Previous studies that consider the perception of a person's turn-taking behaviour and personality traits that are attributed to him are not univocal.

Robinson and Reis [161] conclude from a perception study that interruptors were seen as less sociable and more assertive than individuals who did not interrupt. Goldberg [72] differentiates between power and non-power interruptions and argues that some interruptions are a display of rapport and others of power. This parallels the

distinction between cooperative and competitive speech overlap [74]. Yet a generally cooperative stance does not exclude a competitive interruption. Interruptions by police interviewers are the topic of studies because of the impact they could have on the experience of the suspect or witness [96]. We are not aware of studies that focus on the suspect's turn-taking behaviour related to stance taking and interview strategy. In this section we will show results of our explorative study about how turn-taking behaviour in police interviews is related to the suspect's stance.

3.6.1 Classification of Turn-taking Behaviour

Patterns in turn-taking become visible when we look at speaker transitions through vocal analysis: an acoustic silence paradigm analysing quantitative chronometrical data on something (speech) and nothing (silence between speech) [64]. In two party conversation variations in the vocal activity (speech or silence) of both speakers result in four possible dialogue states: self-speaking, other speaking, none speaking, and both speaking [85]. Transitions between dialogue states create an interaction pattern. Heldner and Edlund [85] distinguish two different classes of silence: gap, a silence in which a speaker transition occurs and $pause$, a silence between two consecutive utterances of one and the same speaker. If more than one speaker is speaking there is overlapping speech, distinguishable in the different classes: $boundary - overlap$, an occurrence of overlapping speech where a speaker transition takes place and $within - overlap$, an occurrence of overlapping speech present during one continuous speech activity of one speaker.

The definitions described above are comparable to, though slightly different from the definitions in Sacks et al. [165], where a pause is a hypernym for silence, silence after a possible point of completion is a gap, and an extended silence at a transition relevant place is a lapse. We adhere to the terminology used by Heldner and Edlund [85]. However, because of the clear difference in and influence of the role of the interlocutors we look at interactions from a third person view and use $police$ and $suspect$ to refer to the active speakers (see Figure 3.5).

Statistics about the occurrences of interaction patterns are useful to find some global characteristics of the type of verbal interaction, but the pattern does not say much about the meaning. We miss the words and the non-verbal communicative signals that contribute to understanding the meaning carried within these interactions. Occurrences of similar vocal activity patterns may carry different meanings. Also, overlapping talk can have different flavours and instead of considering silence as just the absence of talk we can look at silence from a socio-pragmatic point of view.

3.6.1.1 Overlapping talk

Overlapping talk is either competitive, neutral, or collaborative. A competitive overlap is considered indicative for power, control or dominance, or an expression of indifference, aggressiveness, or hostility. Competitive overlap is manifested with high pitch and intensity. A collaborative overlap conveys rapport and is an indicator for coordination and alignment [74]. Schegloff [169] defines four classes of overlapping speech: 1) cooperative overlap such as assisting by completion, 2) non-problematic

Figure 3.5: Illustration of original terminology by Heldner and Edlund [85] and how gaps, pauses, between-speaker overlaps and within-speaker overlaps are classified using observable vocal activity and the dialogue state of the two speakers (police officer and suspect) in a conversation. Additionally, the dialogue state from the third-person view is depicted.

overlapping speech such as chorus, 3) interrupt where the speaker did not finish the utterance and did not yield the floor and 4) back-channel and short feedback, not intended to gain the floor. Overlapping speech of the interrupt class is considered problematic and needs resolution. The overlap can occur after a gap when it is not clear who has the turn; both want to take the floor or when the listener wants to take over the speaker role of the active speaker. The question is, who gives up the fight for the floor? We suggest that stances of interlocutors are a mediating factor here.

3.6.1.2 Silence

Silence can convey meaning and is considered communicative when silence occurs where the rules dictate to speak and the silence is by choice of the speaker. Ephratt [64] defines these silences as eloquent silence: a silence as a means chosen by the speaker with a significant communicative meaning. Verschueren [203] distinguished several causes of a participant remaining silent; causes that we can categorize into two groups:

1) the speaker is temporally disinclined to speak; the speaker is concealing something; the speaker does not have anything to say.

2) the speaker is unable to decide what to say next; the speaker is unable to speak because of strong emotions such as amazement or grief; the speaker has forgotten what he was going to say; the speaker is silent because others are talking.

The first group of causes are intentional according to Ephratt [64]. The second group are considered causes of non-intentional silences from psychological inhibition [113]. We suggest that, as for overlapping talk, stance is a mediating factor for interpreting the semantics of the silence.

3.6.2 Suspect's Stances in the Example Conversation

The interpersonal stance of the suspect, as annotated by multiple annotators, changes during the course of the conversation. We distinguish a global pattern of five segments in the stance of the suspect in Figure 3.6: A) predominantly *defiant*, B) variation between *defiant*, *dependent* and *cooperative*, C) mainly *cooperative*, D) predominantly *defiant* and E) a final *cooperative* moment. These segments are marked in Figure 3.6 showing the suspect's stance as it is annotated by a majority of the annotators. The occurrences of silences and overlaps *gap*, *pause*, $overlap_W$ and $overlap_B$ are shown in Figure 3.6 as well.

To discuss our findings we collected a number of samples showing silence and overlapping speech from the video recordings of our interview. They have been transcribed according to the Jefferson convention[2]. We will provide English translations when we discuss the fragments (in the next subsection). Non-verbal behaviour essential for understanding what is going on is marked in the sample transcriptions. The samples also give the interpersonal stances of the speakers. Analysis of these samples show that certain turn-taking behaviour is related to stance.

[2]e.g. http://homepages.lboro.ac.uk/ ssca1/trans4b.htm, visited 11-12-2015 ([94][128])

The suspect in our interview is Ms. Wassink. She was brought in because her neighbour filed a criminal complaint for assault. Apparently Ms. Wassink became physical after she and the neighbour got into an argument in front of their houses. The police officer read the files and invited her for an interview.

The police officer welcomes the suspect and then explains the goal of the interview in segment A (we refer to the segments from Figure 3.6). The suspect is quiet and withdrawn, resulting in a monologue of the police officer with a number of *pauses* between consecutive utterances.

In segment B the police officer asks concrete questions about the suspect's home situation, the suspect provides (minimal) responses. *Gaps* are frequently observed and *pauses* between suspect utterances appear. The silences are strategically used by the police officer to encourage the suspect to speak.

When discussing the topic of the neighbourhood and the suspect's relation with the neighbours in segment C, the suspect is more talkative and *pauses* between consecutive utterances of the suspect occur frequently. Turn-taking seems to proceed without many problems. During the course of segment C the police officer introduces a new topic; pets. The conversation changes to a casual conversation style; suspect and police officer share their personal attitude towards pets. Both interlocutors are talkative, speak without being addressed by questioning, resulting in an increase of overlapping speech. Both interlocutors do understand what is said by the other. The overlapping speech does not hinder the conversation. The conversation evolves back to an interview type after the police officer initiates the new topic of yesterday's events. The suspect's willingness to talk decreases and *gaps* occur more frequently.

At the start of segment D the suspect explicitly questions the relevance of the proposed question and topic. The suspect provides minimal responses and the frequency of occurrences of *pause* increases. At the very end the new topic of the argument with the neighbour is initiated by the officer. *Boundary-overlaps* occur followed by sequential *pauses* in the speech of the suspect possibly indicating the suspect claims the floor.

In segment E the suspect is talkative and re-selects self as next speaker repeatedly causing *pauses* between consecutive utterances of the suspect. The contributions of the police officer are all *backchannels*.

We see that the topic of the conversation is a factor that influences the stance of the suspect, in particular, if the topic is related to the case at hand. The police officer initiates new topics. The topic influences the talkativity of the suspect and the turn-taking behaviour of both interlocutors. If the suspect is less talkative silences are more frequent. Overlapping speech is mainly present during segment C in which the conversation turns into casual talk. The topic is innocent and non-threatening. The strategy employed is relational, of the sort that Giebels [66, 67] calls 'being kind', stressing shared experiences between police officer and suspect.

3.6.3 How Stance Mediates the Meaning of Silence and Overlapping Speech

The relation between power and the decision to start speaking and continue speaking is visible in samples 3.1, 3.2, and 3.3. Samples 3.1 and 3.2 take place at the boundary of segments B and C. In both fragments the stance of the suspect is *positive* and

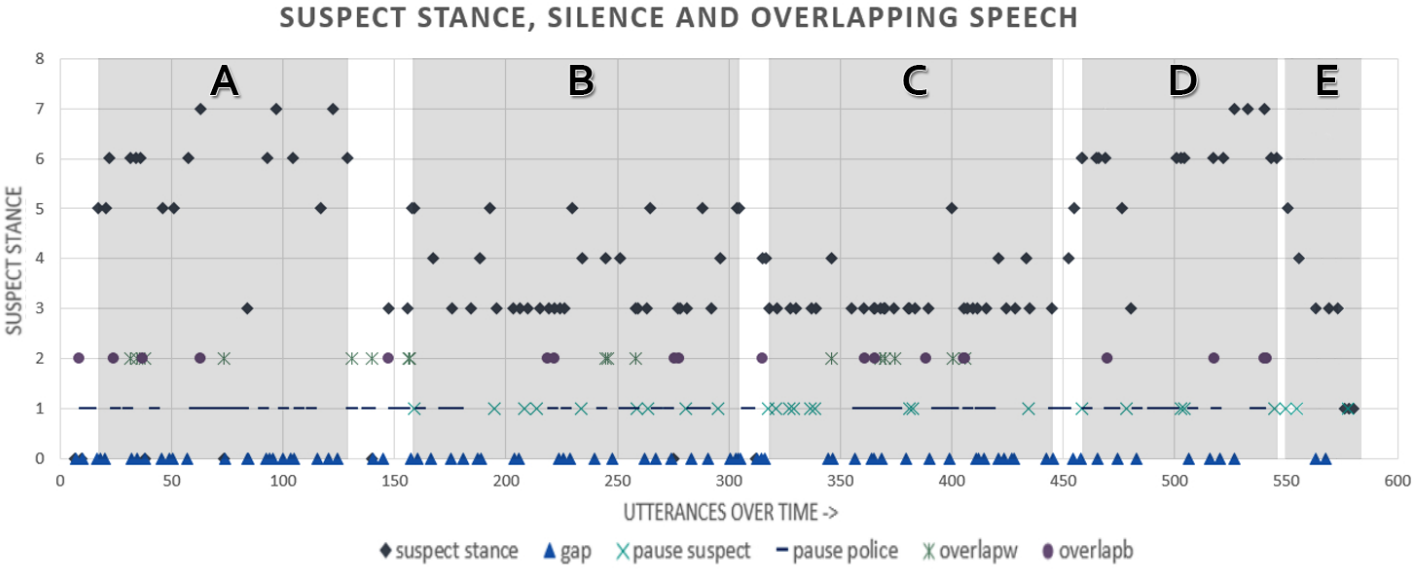

Figure 3.6: Segmentation of the conversation based on suspect's stance (based on an MVA) and the occurrences of silence and overlapping speech.

submissive while discussing the topic neighbourhood.

In Sample 3.1 the officer asks a question addressing the suspect (line 660: *What type of neighbourhood is it?*). After a moment of silence (duration 0.9) the suspect initiates a response (line 665: *yeah*). The officer decides to rephrase a more concrete question and re-selects self, initiating overlapping speech by a short delay in onset (line 669: *A friendly neighbourhood, or not?*). The suspect immediately stops speaking and yields the floor to the officer, resulting in a *boundary-overlap* interaction.

```
660  Police:   SB  wat voor buurt is het?
                   (0.9)
665  Suspect:  NN  nj[ aa ]
669  Police:   SB     ge  zellige buurt, of juist niet?
                   (.)
665  Suspect:  SO  ja vink wel
```

Sample 3.1: Occurrence of post-continue overlap where the initiator (police) wins the floor resulting in boundary-overlap—sample taken from Wassink starting at 00:04:33.540.

In Sample 3.2 the suspect has the turn but decides to stop speaking before sentence completion (line 816: *Yeah, well then they stay in their houses alone with their dog but na yeah then eh*). After a fairly long silence (duration 1.18) the officer selects self as next speaker. Shortly after onset the suspect continues the previous speech act initiating overlapping speech (line 825: *those kinds of things*). The suspect stops speaking causing *within-overlap* interaction. The police officer extends the turn with a second sentence but stops speaking before sentence completion (end of line 821: *that you like to ...*). After a short *pause* (duration 0.27) the officer re-selects self (line 830: *...do things together with other people*). The officer stops speaking at a point of possible completion where the suspect selects self and almost seamlessly starts speaking, taking up the officer's suggestion (line 835: *yeah I like that together with people, yeah*).

```
816  Suspect:  SO  ja gewoon helemaal eeh dan in hun huis blijven
                   zitten enzo met hun hond alleen maar ja naja
                   dan eh
                   (1.18)
821  Police:   SB  ieder voor zich, god voor ons allen en jij
                   zegt van, geeft eigenlijk een [beetje aan van
                   dat je] toch wel een gemeenschapsmens bent.
                   dat je dat graag
825  Suspect:  SO  [van die dingen]
                   (0.27)
830  Police:   SB  met andere mensen iets samen wilt doen
                   (.)
835  Suspect:  SO  ja ik vind wel leuk memensen samen, ja
```

Sample 3.2: Occurrence of post-continue overlap where the initiator (suspect) loses the battle for the floor resulting in within-overlap—sample taken from Wassink starting at 00:05:39.150.

Sample 3.3 takes place in segment C where the stance of the suspect is fully *co-*

operative. While discussing the topic pets, the conversation type devolved to casual conversation —losing the interviewer and interviewee roles. After a short comment by the officer (line 888: *How nice*) the suspect selects self as next speaker (line 893: *Yes, my mum in particular*). The police officer re-selects self as next speaker initiating overlapping speech a brief moment after onset. The suspect continues speaking resulting in a *within-overlap*. However, the suspect does not complete the sentence, pauses and takes up what was said by the officer during the overlapping speech.

```
888  Police:   SB  wat leuk
                   (.)
893  Suspect:  SO  ja[ah mn moeder hoor vooral]ik dr
897  Police:   SB    [ik heb dr vier dr vier ]
                   (0.64)
902  Suspect:  SO  ja echt?
```

Sample 3.3: Occurrence of overlapping speech—sample taken from Wassink starting at 00:06:07.500.

Samples 3.1 and 3.2 show that the interlocutor with higher power is more likely to win a battle for the floor. This results in a *within-overlap* when the suspect initiates the overlapping speech and in a *boundary-overlap* when the police officer initiates the overlapping speech. The examples also illustrate (our hypothesis) that the interlocutor with higher power is more likely to self-select as next speaker during a silence after an incomplete utterance of the other. Sample 3.3 illustrates that even a slight change in power—resulting from a change in conversation type—increases the likelihood for a suspect to self-select as next speaker and continue speaking during a battle for the floor.

The relation between affiliation and the pragmatic interpretation of silent responses is illustrated by samples 3.4, 3.5, and 3.6. Sample 3.4 takes place in segment B where the stance of the suspect is submissive and tends to be positive. The officer asks a check question addressing the suspect (line 347: *I've understood that you live in the Broekstraat*). The answer is presented by a non-verbal affirmation in the form of a head nod.

```
347  Police:  SB  ik heb begrepen dat jij aan de Broekstraat woont
                  ((looks up at suspect))
                  (0.73) ((suspect headnod))
352           SB  ja?
                  (.)
357           SB  woon je daar alleen of met iemand anders?=
```

Sample 3.4: Occurrence of pause where participation of the other interlocutor takes place by non-verbal behaviour—fragment taken from Wassink starting at 00:02:23.406.

Sample 3.5 takes place in segment A when the stance of the suspect is submissive and hostile. *Pause* in the speech of the police officer occurs frequently and sequentially between incomplete utterances of the police officer. In lines 206-216 (*Your name is Sabrina. Sabrina Wassink I've gathered. I don't know you. You don't know me either*) the officer wants to check some data from his sheet, keeping an eye on the suspect to

read her response when he pauses waiting for confirmation. The suspect's responses are non-verbal and minimal. She doesn't feel much like getting to know one other as the officer proposes.

```
191  Police:  NN  ehm
                  (0.69)
196           NN  maar
                  (0.4) ((suspect looks away))
201           NN  goed, ik eheh h
                  (0.63)
206           SB  je heet Sabrina
                  (0.77) ((suspect head nod))
211           SB  Sabrina Wassink heb ik begrepen
                  (0.46) ((suspect head nod))
216           SB  ik ken jou verder niet
                  (0.58) ((suspect head shake))
221           SB  jij kent mij ook niet
```

Sample 3.5: Occurrence of sequential pauses filled with a non-verbal response if the utterance was a polar question—sample taken from Wassink starting at 00:01:14.913.

Sample 3.6 takes place in segment D when the global stance of the suspect is hostile and predominantly submissive. The police officer asks a polar question (line 1235: *Did anything out of the ordinary happen after that?*). The suspect provides a non-verbal negative response. The police officer elaborates on the question several times (line 1243: *or after you returned from Zwolle?*; line 1253: *also not in the neighbourhood?*; line 1258: *Did you bump into someone with a dog?*) thereby indicating that he has some specific information he wants to check. The suspect responds repeatedly with head shakes and incertitude facial expressions but after some time she starts speaking, admitting in a reluctant way that she may have seen the dog (line 1263: *yeah, eh might be but hhh*).

```
1235  Police:   SB  is dr nog iets bijzonders gebeurd daarna?=
1239            SB  =###########= ((suspect head shake))
1243            SB  =of nadat je van Zwolle teruggekomen bent?
                    ((suspect head shake))
                    (3.12) ((suspect head shake))
1248            NN  hmm
                    (2.2)
1253            SB  ook nie in de buurt?
                    (1.51) ((suspect incertitude expression))
1258            SB  nog iemand tegen gekomen met een hondje?
                    (0.96)
1263  Suspect:  TO  ja eh vast wel maa hhh
```

Sample 3.6: Occurrence of sequential pauses while the suspect aims to conceal information—sample taken from Wassink starting at 00:08:05.616.

Sample 3.4 shows an example of a cooperative suspect who provides a response in a non-verbal way nodding her head. This silence could be interpreted as *have*

nothing to say. In Sample 3.5 the suspect provides a complete non-verbal response but given the more hostile stance of the suspect and other non-verbal behaviour (i.e. looking away from the speaker) the silence can be interpreted as the suspect feeling *disinclined to speak*. In Sample 3.6 the withdrawn stance of the suspect indicates that the absence of speech is aimed at *concealing* possibly incriminating information.

We have seen that stance of the suspect on the affiliation axis shows a correlation with interpretation of silence by the suspect. A silent but contributing response is related to either timidness (positive stance) or withdrawal (hostile stance). A silent response intended to withhold information is only observed in relation to a hostile stance. The global stance on the power axis shows a correlation with talkativeness. The global stance of the suspect is predominantly submissive (low power). This submissiveness is related to: 1) the decision to speak only when selected as next speaker by the other interlocutor and 2) the decision to yield the floor during overlapping speech independently of the initiator and onset of overlap. This difference in speaker activity reduces when the power levels of the suspect and the police officer become more equal.

3.7 Conclusions

Based on the outcome of our reliability analysis we are convinced that Leary's theoretical model makes sense as a framework for analysing and describing the interactional stance that people take towards each other in a social encounter. Leary's Rose provides terminology to come to a reasonable agreement between subjects about 'what is going on' in a police interview in terms of stance taking and the dynamics of the behaviours and the effects they have on turn-taking by the interlocutors. Sometimes, it turns out to be hard for outside observers to tell what the stance of the participant is. The multi-flavoured expression of stance in general seems to be an important cause of disagreement between judges. Another cause is the forced choice annotation procedure and the fuzzy character of the meaning of the words used. Judges were forced to make a choice where it is often hard to make a choice. Analysis of the annotated corpus has shown that it is indeed the case that when annotators disagree in their choice they choose labels that are next to each other in Leary's Rose. Moreover, they agree about the direction in the two main dimensions. If we use a majority voting meta-annotator the inter-subjective content of Leary's view on stance is clearly revealed in the changes in stance. Although judges often do not agree about the exact label on the level of speaker segments they do agree on the global dynamics of the stance changes during a police interview.

The general lesson we learned from this is something we already knew before but sometimes forget: do not use too precise measures for fuzzy phenomena. Based on our analysis of the annotations the annotation instruction can be improved, in particular regarding back-channels and short feedbacks.

This explorative study into the relation between a suspect's stance and the types of overlaps, interruptions, and silences indicates that the interview topic and in particular how the topic is related to the case at hand is an important factor that influences the stances taken by the subject. Stances and roles seem to be mediating factors for

the meaning of overlaps and silences in suspect interviews. With this we could try to face the challenge of incorporating these findings into a computational model of a suspect character so that it simulates believable behaviour that expresses the suspect's stance as a response to the learner's strategy and stance taking. For this, we analysed more police interviews to substantiate our preliminary findings and to build better models for different suspect characters in chapter 5. But first, in the DPIT corpus professional actors play the role of suspect. This raises the question, are acted stances a viable substitute for 'real' stances? In the next chapter (4) we will investigate this by looking at the effect actor proficiency has on the recognisability of the portrayed stance behaviour.

4

The Recognition of Acted Interpersonal Stance in Police Interrogations

In this chapter[a] we will report on judgement studies regarding the perception of interpersonal stances taken by humans playing the role of a suspect in a police interrogation setting. The main question we ask in this chapter is: do human judges agree on the way they perceive the various aspects of stance taking, such as friendliness and dominance and what is the influence of actor proficiency? Four types of stances were acted by eight amateur actors. Short recordings were shown in an online survey to subjects who were asked to describe them by selecting appropriate adjectives from a list of adjectives. Results of this annotation task will be reported in this chapter. We will explain how we computed the inter-rater agreement with Krippendorff's alpha statistics using a set distance metric based on theory. Results show that for some of the stance types observers agreed more than for others. We further investigated the effect the expertise of actors has on the perception of the stance that is acted. We compared the fragments from amateur actors to fragments from professional actors taken from popular TV-shows. Some actors are better than others, but validity (recognizing the intended stance) and inter-rater agreement do not always go hand in hand.

[a] Based on *Bruijnes, M., op den Akker, R., Spitters, S., Sanders, M., & Fu, Q. (2015). The recognition of acted interpersonal stance in police interrogations and the influence of actor proficiency. Journal on Multimodal User Interfaces, 1-24.*[33, 178]

4.1 Introduction

People quickly form impressions of each other's personality and interpersonal stance (attitude). This also holds when people encounter virtual humans [38]. Research has also shown that when several raters were asked to encode interpersonal dispositions of people the agreement was low [68]. Nevertheless, if we build realistic and believable virtual suspect characters we need to pay attention to the relation between observable nonverbal behaviours and the way the police trainee perceives and judges the character and attitude of the virtual suspect. Moreover these characters have to be interpretable in the sense that "the user must be able to interpret their responses to situations, including their dynamic cognitive and emotional state, using the same verbal and nonverbal cues that people use to understand one another" [103]. A virtual human does not have the same cognitive capabilities as a human. This makes an interaction between them what Baranyi and Csapó [14] call an *inter-cognitive communication*. The challenge here is that a virtual human has to give the human the impression it has cognitive capabilities that are similar to that of the human. When successful their interaction should be called an *intra-cognitive communication*.

Basically there are three different methods to follow when building models for the generation of the dynamic behaviours of virtual humans. In the artist method, virtual characters are created and their behaviours and expressions are generated based on intuition after which we observe how the character is perceived. Another term that is used in the literature for the artist method is puppeteering [147]. Contrary to the artist method the analytical approaches are towards finding generalizable rules or statistics about the typical behaviours that express a certain stance. In the design method different designers are given a virtual human and a set of basic behaviours and expressions. For a number of stances, the designers are asked to generate behaviours and expressions with the virtual human that they believe express the given stances. The results are analysed to see how often designers used each of the basic behaviours for each of the stances. The statistical behavioural model is used to generate the most likely (combination and sequences of) behaviours when the virtual characters takes a certain stance. This method was followed by Chollet et al. [45]. In this chapter we will report on a second analytical method which is based on the analysis of stances played by human actors. The scene of play is that of a face-to-face police interview where one police officer interrogates a suspect. In a number of judgement surveys recordings of human actors were presented to human judges who were asked to label the stance expressed by the actors in the fragments of the interviews. The method raises the following three issues.

A) The collection and selection of the audio/video fragments that show the behaviours. Do we use actors and how do we generate specific stance behaviours in lab settings, or do we use real-life recordings?

B) The task of the human judges that label the data. What is the annotation procedure, the label set, is it categorical or continuous?

C) The way reliability of the labelled data is measured. Can we assume ground truth? How to compute inter-rater agreement?

A) We used two different types of audio/video recordings: for the first experiment we had non-professional actors play a specified stance in a given interrogation scenario or we asked them to respond to a stance taken by the interviewer in a given scenario. For a second experiment we chose a number of fragments from TV series, showing professional actors playing the role of a suspect. B) We will report about different ways of labelling stance. In the first experiment we used a semi-free annotation format where raters could choose from a given set of adjectives that describe the stance shown. In the second experiment we used a three-dimensional continuous annotation schema consisting of three 5-point Likert scales for dominance, affiliation, and spontaneity. C) Several methods have been used to measure the validity and reliability of the labelled data. Can we assume ground truth about the stance that actors portray?

With the analysis of acted social, emotional, or stance behaviour of human actors comes the consideration of how natural this acted behaviour is. The question is, can an actor show behaviour on demand of the researcher and how close to real-life behaviour is this on-demand behaviour. Bänzinger and Scherer [13] distinguish three categories of 'naturalness' of recorded behaviours: (1) Natural behaviour occurs in real-life settings and is not directly influenced or controlled by the researcher. (2) Behaviour can be induced in a controlled (laboratory) setting that is designed to elicit the behaviour in which the researcher is interested. (3) Portrayals of behaviour by actors upon instruction by the researchers. It is common for elements from these categories to co-occur, for example induce emotions with the instruction of the to be portrayed behaviour. The underlying feelings and emotions of 'natural' behaviour cannot be directly assessed, they can only be inferred from observations or post-hoc reports by the 'actor'. This is also a problem when using TV clips from professional actors as we have done in this study. These clips have the advantage of showing more natural behaviour than that of amateur actors, but the intent of the behaviour cannot be validated. The underlying emotional 'intent' is available from induced behaviour or portrayals of behaviour as it is part of the instruction or design of the setting that induced the behaviour: 'if you ask for a smile, you get a smile'. In this study we have therefore also asked (less experienced) actors to act out behaviours based on the researcher's instructions. By knowing the actual intent of the acted behaviour, we can investigate whether others perceive the behaviour as intended. Obtaining variations of a behaviour or multiple instances of a behaviour from the same actor can be difficult or impossible in 'induced' or 'natural' behaviour. When asking actors to portray behaviour it is possible to get all variables of the behaviour from each actor. This allows comparison of the same behaviour from different actors. In addition, Busso and Narayanan [37] took a "deeper look at the current settings used in the recording of the existing corpora [which] reveals that a key problem may not be the use of actors itself, but the ad-hoc elicitation method used in the recording." They suggest that portrayals will be as close to natural behaviour as possible if care is taken to: (1) contextualize the (social) setting properly; (2) combine the acting styles 'scripted' and 'improvisation' to have both the influence the researcher needs and the freedom the actor needs in their emotional expression; (3) give actors the time to prepare or rehearse their acts, or use skilled actors if possible; and (4) define the references used

to describe the emotional and social acts as they are often blurred and partial to subjectivity [13]. An actor might know, because of his or her training, which behaviour will be perceived as the intended behaviour and consequently show this behaviour. This is troublesome for those behaviours where there is a discrepancy between typical classifications by observers and what people (not acting) show in real-life situations. For example, Strömwall et al. [182] show that there is a difference between what people believe to be indicative for deception and what is actually indicative. For this reason, an actor who is instructed to show deceptive behaviour can unknowingly show unnatural behaviour; behaviour that a deceiving person would not show. And an observer might rate this behaviour as deceptive behaviour.

Acted portrayals were suggested as unsuited for applied research purposes. Taking a different stance, Bänziger et al. [12] argue that portrayals produced by appropriately instructed actors are analogue to expressions that do occur in selected real-life contexts. Also, acted portrayals - as opposed to induced or real-life sampled emotional expressions - display the most expressive variability and therefore constitute excellent material for the systematic study of nonverbal communication of emotions. “In everyday life”, they argue, “emotional expressions are directed to receivers with different degrees of intentionality. Some expressions might be truly ‘spontaneous’, not directed or intentionally regulated to have an impact on a receiver; whereas acted portrayals are by definition produced intentionally and directed to a receiver.” In our second experiment we asked judges to score the spontaneity of the actor’s stance.

In section 4.2 we will discuss stance taking in the interesting context of the police interview in which both parties are often not on the same wavelength. In section 4.3 we will discuss related work. In section 4.4 we will explain the method of our study and in section 4.5 the outcomes. In section 4.6 we will investigate the effect the expertise of the actor has on the perception of his behaviour. We will draw conclusions in section 4.7.

4.2 Interpersonal Stance

Inbau et al. [92] list five essential principles that must be followed by the interviewer in order to decrease the probability of making erroneous inferences from a suspect’s behaviour.

1. There are no unique behaviours associated with truthfulness or deception.
2. Evaluate the consistency between all three channels of communication: verbal, paralinguistic (among which response timing and length) and nonverbal.
3. Evaluate paralinguistic and nonverbal behaviours in context with the subject’s verbal message.
4. Evaluate the preponderance of behaviour occurring throughout the interview.
5. Establish the suspect’s normal behavioural patterns.

The technical notion of ‘nonverbal behaviour’ refers to a range of observable aspects of communicative phenomena in human encounters. It includes turn-taking,

the switching of speaker and listener roles, prosodic aspects of speech, body postures and facial expressions. All these aspects may reflect stance taking. Here we report on the study of the typical postures and facial expressions that signal the stances. Op den Akker et al. [141] reported on the relation between stance taking and turn-taking in police interviews. Deception and lying happen in police interviews more than in many other encounters and interviews. Suspects make statements to make the other believe something that is not true. A lie is not a special type of sentence with it own linguistic or para-linguistic identifiers. Similarly, stances that are taken deliberately to make some impression on the other are not distinguishable from 'sincerely' taken stances. A deliberate and 'strategically employed' stance does not have special types of observable features that differ from 'sincere' stances. So despite that people often purposely take a stance in the context of police interviews, we do not expect that the relation between stances and their observable behavioural features is different from that in other contexts where people act more spontaneously.

Research has demonstrated the value of the circumplex model for integrating a broad range of psychological topics. Researchers have mostly used Wiggins' English Interpersonal Adjective Scales (IAS) from [212]. Rouckhout and Schacht found a circumplex structure underlying a comprehensive set of Dutch interpersonal adjectives [164]. The configuration was divided into eight segments isomorphic to the IAS octants. Fifteen adjectives from each segment were used to form eight preliminary Dutch interpersonal scales. Table 4.1 shows the Dutch adjective scales (Table 4.2 shows the English translations) from [164]. A representative random selection of the adjectives, highlighted in blue, was used in our study, see sections 4.4 and 4.5.

Table 4.1: Dutch adjectives scales for the categories of the interpersonal circumplex (the English translations can be found in Table 4.2). From: [164].

DH		SH		SP		DP	
Concurrerend	**Aggression**	**Opstandig**	**Teruggetrokken**	**Depend**	**Cooperative**	**Helping**	**Leidend**
Eigenwijs	Onbarmhartig	Afhankelijk	Verlegen	Pretentieloos	Minzaam	Ongedwongen	Onwankelbaar
Cynisch	Geniepig	Onbeholpen	Onderdanig	Eenvoudig	Lief	Charmant	Extravert
Gehaaid	Pretentieus	Gezagsloos	Schuchter	Moederlijk	Tolerant	Mededeelzaam	Praatgraag
Aanvallend	Wreedaardig	Tactloos	Stil	Braaf	Liefdevol	Vrolijk	Krachtig
Vrijpostig	Grof	Onpersoonlijk	Beschaamd	Doodernstig	Inschikkelijk	Spontaan	Geestdriftig
Autoritair	Bevooroordeeld	Kortaf	Gereserveerd	Nederig	Zachtmoedig	Opgewekt	Geraffineerd
Dominant	Doortrapt	Asociaal	Twijfelend	Onschuldig	Barmhartig	Galant	Hardnekkig
Vechtlustig	Brutaal	Egocentrisch	Timide	Overbezorgd	Bedaard	Attent	Fier
Impulsief	Achterdochtig	Liefdeloos	Introvert	Buigzaam	Gewillig	Voorkomend	Resoluut
Onstuimig	Tegendraads	Onoprecht	Terughoudend	Weekhartig	Onbevooroordeeld	Sociaal	Ambitieus
Ongegeneerd	Sluw	Bedrieglijk	Gesloten	Alledaags	Bescheiden	Vriendelijk	Spraakzaam
Dikhuidig	Geslepen	Oneerbiedig	Schuw	Onopvallend	Teerhartig	Loyaal	Kordaat
Onbedeesd	Listig	Arrogant	Gemaakt	Bedeesd	Zachtzinnig	Tactvol	Vol vuur
Onverlegen	Despotisch	Afgunstig	Afzijdig	Volgzaam	Plooibaar	Menslievend	Vastbesloten
Onbeschroomd	Schaamteloos	Intolerant	Naïef	Meegaand	Teergevoelig	Behulpzaam	Levendig

4.3 Related Work

In studies on how people communicate with posture, gesture, stance, and movement it is often argued that all movements of the body have meaning (i.e. are not accidental), and that these nonverbal forms of language (or para-language) have a grammar that can be analysed in similar terms to spoken language (e.g. [22]). Birdwhistell [22] estimated that "no more than 30 to 35 percent of the social meaning of a con-

Table 4.2: (Translated from) Dutch adjectives scales for the categories of the interpersonal circumplex (the highlighted words represent the selection used in this experiment). From: [164].

DH		SH		SP		DP	
Compete	**Aggression**	**Defiant**	**Withdrawn**	**Dependent**	**Cooperative**	**Helping**	**Leading**
Cocky	Unmerciful	Depending	Shy	Unpretentious	Affable	Laid-back	Unfaltering
Cynical	Sneaky	Awkward	Submissive	Simple	Sweet	Charming	Extrovert
Dodgy	Pretentious	Authority negligent	Bashful	Maternal	Tolerant	Communicative	Garrulous
Offensive	Cruel	Tactless	Quiet	Good	Loving	Cheerful	Powerful
Bold	Foul	Impersonal	Ashamed	Dead serious	Accommodating	Spontaneous	Enthusiastic
Authoritarian	Biased	Curt	Reserved	Humble	Meek	Excited	Sophisticated
Dominant	Sly	Antisocial	Doubting	Innocent	Merciful	Gallant	Stubborn
Pugnacious	Cheeky	Egocentric	Timid	Overanxious	Composed	Attentive	Proud
Impulsive	Suspicious	Loveless	Introvert	Flexible	Willing	Courteous	Resolute
Impetuous	Rebellious	Insincere	Reserved	Soft-hearted	Unprejudiced	Social	Ambitious
Unabashedly	Cunning	Deceitful	Closed	Casual	Modest	Friendly	Talkative
Thick-skinned	Polished	Irreverent	Shy	Discrete	Tender-hearted	Loyal	Firm
Not timid	Crafty	Arrogant	Artificial	Timid	Gentle	Tactful	Passionate
Not shy	Despotic	Envious	Aloof	Docile	Pliable	Humane	Resolved
Unabashedly	Shameless	Intolerant	Naive	Submissive	Sensitive	Helpful	Lively

versation or an interaction is carried by the words". Communication of interpersonal attitudes is one of the five primary functions of nonverbal behaviour [4]. Mehrabian [129] found that body orientation affects the conversation. Body language comes in clusters of signals and postures, depending on the internal emotions and mental states. Recognizing a whole cluster is thus far more reliable than trying to interpret individual elements. Smith-Hanen [176] reports a study in the perception of empathy through body postures. The author concludes that more attention should be focused on the nonverbal channels of communication in the training of counsellors. Dael et al. [52] adopted the Body Action and Posture (BAP) coding system to examine the types and patterns of body movement that were employed by 10 professional actors to portray a set of 12 emotions. The authors investigated to what extent these expression patterns support explicit or implicit predictions from emotion theories. The study revealed that several patterns of body movement occur systematically in portrayals of specific emotions, allowing emotion differentiation.

Nonverbal behaviour is studied in the literature using two similar techniques. The first method involves the participant viewing videotaped actors performing certain actions and the second method involves having the participant sit in a certain posture and then self-reporting their emotions. The use of actors in studying emotions and stance has a number of advantages over collecting real data. But, are acted stances representative of real stance? Or, does the experimental setting in which stance is generated and the fact that subjects know that the stances are acted reduce the validity of the outcomes? As far as we know there has not been a study of the validity of acted *stances*. In the context of *emotion* research in speech the issue has been considered by Wilting et al. [213]. Based on a perception experiment, they concluded that acted emotions (especially negative ones) were perceived more strongly than the real emotions. The suggestion is that actors do not feel the acted emotion, and may engage in overacting, which casts doubt on the usefulness of actors as a way to study real emotions. Acting has a particularly strong effect on the spoken realization of emotions. Busso and Narayanan [37] argue that if certain conditions are satisfied professional actors can be used for valid emotion research. Conditions for the 'gener-

ation of emotions' are that professional well-instructed actors should be used. Enough context should be given to actors for eliciting the emotional state. Asking actors to read a sentence aloud in a 'sad voice' is not good practice for building an emotion database. Other conditions concern the perception and descriptions of the emotions. An interesting approach –given the goal of our own project– is the sensitive artificial listener (SAL) approach in which emotions are induced in a context of a human interacting with a virtual character/machine [59]. For a recent survey on affective body expression perception and recognition refer to [107]. See Table 4.3 for an overview of some nonverbal behaviour that the literature has shown to be important or occurring more during a stance.

Culture influences the stance people take towards others and the meaning of certain behaviours and expressions [11, 69]. Endrass and André [63] integrated cultural factors into models of virtual characters. Police interview studies have shown how differences between high culture and low culture have impact on how sensitive suspects are for the different interrogation strategies and stances that police officers apply [17].

Table 4.3: Some typical stance behaviours from the literature.

	Dominant	Submissive	Hostile	Friendly
Head movement	Tilt head up, Orient head toward other, Shake head [41]			Tilt head up, Orient head toward other [41]
Hand gesture	Movements directed away [133], High gesture rate while talking [41], Initiate hand shaking [41]	Movements directed inward [133], Object-Adaptor, Self-touch [154]	Self-protection gestures [204], Folding arms [69, 154]	Touch other [154], Object-Adaptors [154], Initiate hand shaking [41]
Posture	Space filling & Asymmetric postures [69, 133], Erected posture [41]	Shrinking postures [69]	Distant postures, Barier postures [69]	Physically close postures, Close interaction or direct orientation [69, 154]
Leg movement	Wide stance of the legs [133]		Rhythmically moving legs [204]	
Facial expression	Facial anger [62], Self-assured expression [154], Expressing face [154]	Facial sadness [62, 154]	Facial disgust [62], Facial anger [62, 154]	Smile [38, 62, 69, 86]
Gaze behavior	More gaze [38, 154], Gaze for a long time [154]	Avert gaze [154]	Gaze for a long time [154]	Mutual gaze [38, 154]
Focus of attention		Pay attention to other [154]		Pay attention to other [154]
Turn taking	Overlapping speech [204]	Pause often [133]	Overlapping speech [204]	
Vocalization	Loud voice [133], High pitch [204], High rhythm [204]	Low voice [133]		

4.4 Method: Generating and Annotating Stances

The method followed in this research consisted of two parts. First of all, clips of the interpersonal stances were generated using actors. This will be discussed in Section

4.4.1. The validity of these depicted stances was assessed by means of annotation. The annotation process will be discussed in Section 4.4.2. The results of this annotation process will be discussed in Section 4.5.

4.4.1 Generating Interpersonal Stances

The clips of the interpersonal stances were generated by using actors. Eight actors took part in this experiment. Four of them were members of a theatre club and thus, had some acting experience. Each actor had to depict four stances. The stances correspond to the quarter segments of Leary's Rose (see Figure 1.1). The four segments are abbreviated as dominant-positive (DP), submissive-positive (SP), submissive-hostile (SH), and dominant-hostile (DH).

All actors were given the same scenario. They had to imagine they were suspected of shoplifting and in the middle of an interrogation. Then, they watched a computer screen where a video fragment shows a police interrogator addressing them and asking them what happened. So, the only thing the interrogator says is: *"Vertel eens, wat is er gebeurd?" ("Tell me, what happened?")*. The actors were then asked to give a short response (max. $10\ s$) with a certain stance. This process was repeated until all four stances were depicted. This produced 32 video recordings. These were used in the survey.

The actors differed in the instructions they got on how to depict interpersonal stances. Half of the actors were selected to the 'theory condition' and the other half to the 'role play condition'. Subjects with and without theatre experience were evenly distributed over both conditions. The actors in the 'theory condition' were given theoretical instructions about Leary's Rose [116]. To help them get an even more concrete idea of what the stances mean, several adjectives that capture the meaning of the stances were given. Using these instructions, the actors had to react to the virtual interrogator according to these stances. The adjectives were a random selection from each category of adjectives used in [164], see Table 4.1. The summary of their instructions were captured in an image, that is shown in Figure 4.1. In this condition we tried to ensure that the actors were influenced as little as possible on how they should react to the interrogator on the screen, because it was important that the reaction should be an interpretation of the stances that came from the actors and not from the researchers. The actors in the 'role play condition' were given a specific scenario for each stance that was directly linked to the interrogation setting and to the question of the interrogator. An example of such a scenario is given in Figure 4.2. The scenario was supposed to provoke a reaction in a certain stance in a more natural way than was the case in the 'theory condition', as the workload of processing and interpreting the theory was reduced and actors could put their resources into entering into the part they were playing. Both conditions were taken into account to see whether actors indeed need the information in a scenario format to put down a good performance or if a theoretical instruction is good enough.

4.4.2 Annotating Interpersonal Stances

In an online survey, participants($n = 84$) were shown video fragments in which an actor displayed a specific intended stance. The participants had to select a number of adjectives from 32 different adjectives that best fitted how they would describe the stance taken by the actor in the video fragment. A convenience sample was used that consisted largely of students. The participants were each asked to annotate 8 fragments from a total of 64, 32 with sound and the same 32 fragments without sound. The distinction between with and without audio was used to check whether people were better at recognizing interpersonal stances when they also heard what was said and how it was voiced. The video fragments were assigned randomly to the participants, but in such a way that a participant viewed exactly one clip of each actor.

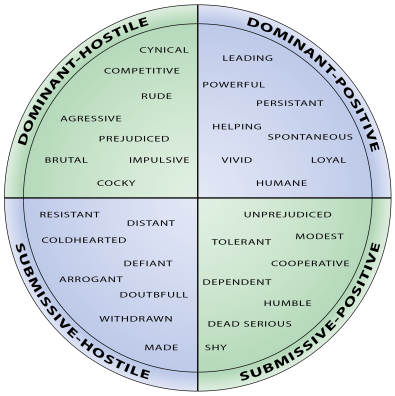

Figure 4.1: Summary of the instructions given to the actors in the 'theory condition' in order to express interpersonal stances.

A semi-forced format was used for annotating the fragments, meaning that participants were given a list of 32 adjectives and were free to select any number of adjectives (with a minimum of four) that they thought fit the stances expressed in the fragments. The list of adjectives was the same as used in the theoretical instruction for the actors. Furthermore, the list was presented in a random order to prevent that the first category would be over represented in the data because of order effects. A screenshot of the survey is shown in Figure 4.3.

In a forced choice format for stance recognition subjects had to select one label or word describing the stance that was shown. The format was debated and some authors advocated a free choice format where subjects were free to choose their own wording to describe the observed stance. Limbrecht-Ecklundt et al. [119] discussed the pros and cons of both formats. To overcome the problems with forced choice we use a semi-forced format. It has the disadvantage that it is less straightforward to compute accuracy and inter-rater agreement.

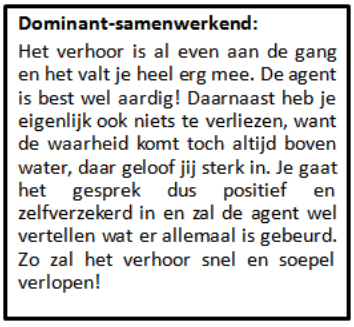

Figure 4.2: Example of a scenario given to the actors in the 'role play condition' to help them express interpersonal stances. This is the scenario for the stance DP. Translation: ***Dominant-positive:*** *The interview is well under way and it isn't that bad. The officer is actually nice! Besides that, you don't have anything to lose because the truth will come out eventually, you strongly believe in that. You approach the conversation positively and with confidence and you will explain to the officer what has happened. That way the interview will be quick and smooth!*

4.5 Results: Recognizing Stance

The clips were acted and the question is whether the acted stances are acted in such a way that observers recognize the portrayed stance. First, we will focus on the distributions of the responses to get a first indication of how well people performed at annotating the videos. Second, the individual judgements will be investigated to see how well individuals recognized the stances. We computed inter-rater agreement. The best recognized videos for each stance were annotated to extract key poses, gestures and facial expressions that can be used when a conversational agent has to convey a certain stance. Note that when validity is high inter-rater agreement is also high, but a high inter-rater agreement does not imply that the agreed stance coincides with the intended stance; validity can be low for some or all stances.

4.5.1 Distribution of Responses

To get a first indication of how well acted stances were recognized, we tested whether adjectives belonging to the depicted stance were chosen more often than adjectives from other stances. To adjust for respondents choosing many adjectives when annotating a fragment and therefore having a bigger influence, calculations have been made for each annotation reporting the percentage of adjectives that belong to the different stance categories. The distributions of these percentages will be used in this section.

For each of the four stances that were depicted by the actors, a pie chart was made that shows the mean percentages of annotated adjectives belonging to each stance-category. These pie charts can be found in Figure 4.4. The figure gives a first indication of how good respondents were at recognizing the depicted stances. It is striking that stance category SH was chosen the most by the respondents independent

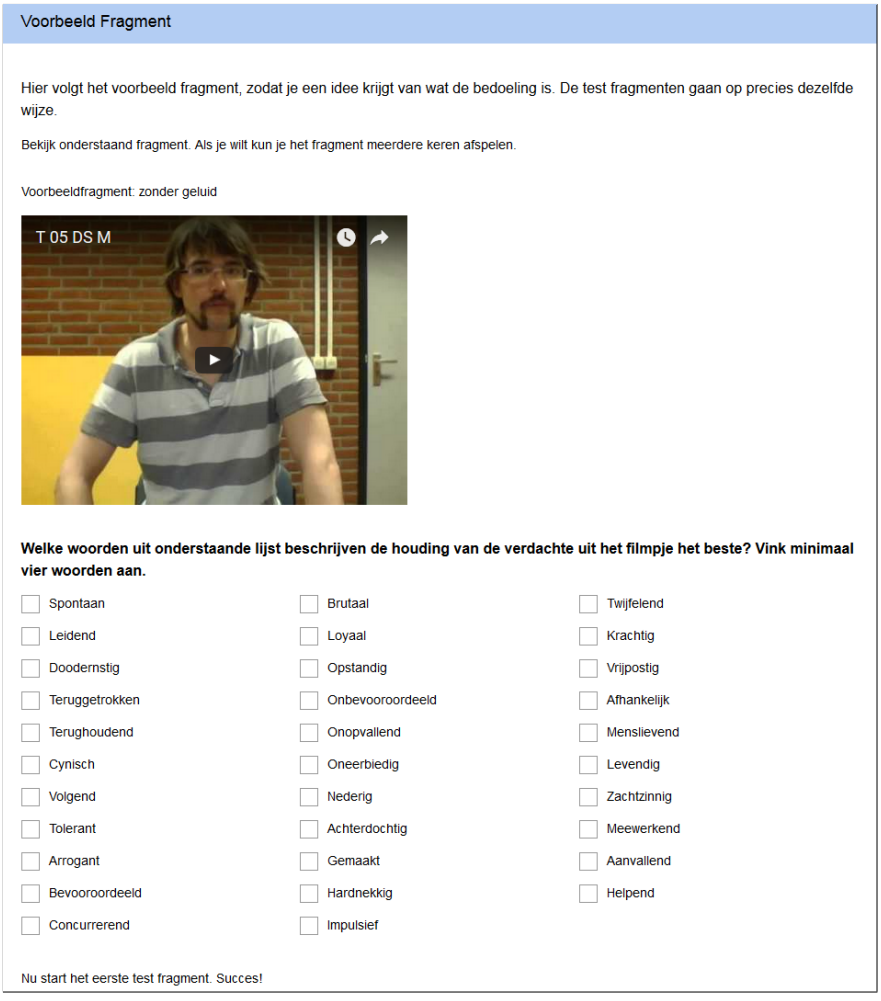

Figure 4.3: Screenshot of the survey used to detect how well people recognized the acted stances.

of what stance the actor depicted. The next step was to test whether the differences between chosen stance categories that seem apparent in the pie chart are significant.

We performed the Kolmogorov-Smirnov test to test the distribution of the data and concluded that for each acted stance category the data was not normally distributed (all $p < .001$). Therefore, the Kruskal-Wallis test was used to test whether the chosen stance categories differed. Mean scores, standard deviations and test values are shown in Table 4.4. It can be seen here that for all the acted stance categories there were differences in the percentage annotated adjectives between the chosen stance categories (all $p < .001$). Next, we investigated where these differences were.

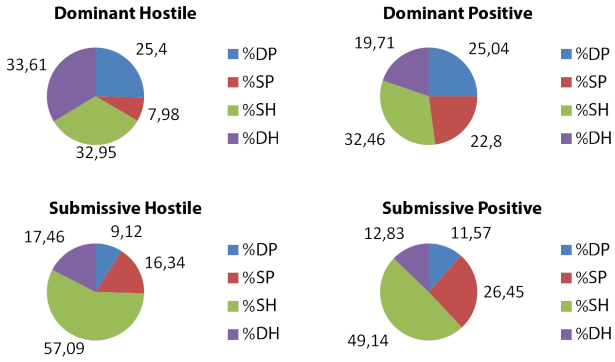

Figure 4.4: For each of the acted stances, the pie chart shows the mean percentages of chosen adjectives belonging to the four stance categories.

Table 4.4: For all the stances that respondents could choose from, the mean probabilities, standard deviations and number of observations given a certain acted stance are shown. The table also shows χ^2- and p-values that test if these mean probabilities differ. The probabilities represent the percentages of annotated adjectives belonging to a certain stance

	Chosen stance												Testvalues	
	DP			DH			SP			SH				
Acted st.	mean	sd	N	mean	sd	N	mean	sd	N	mean	sd	N	χ^2	p
DP	0.250	0.244	162	0.197	0.216	162	0.228	0.236	162	0.325	0.254	162	21.4	$< .001$
SP	0.116	0.179	157	0.128	0.173	157	0.265	0.246	157	0.491	0.247	157	193.8	$< .001$
SH	0.091	0.160	175	0.175	0.207	175	0.163	0.224	175	0.571	0.258	175	274.8	$< .001$
DH	0.255	0.224	178	0.336	0.211	178	0.080	0.148	178	0.329	0.243	178	163.4	$< .001$

To find, for each acted stance category, where the differences between chosen stance categories were, multiple Wilcoxon rank-sum tests were done. It was expected that the stance category that was represented the most would accord with the acted stance category. Therefore we only tested if this category differed from the other three stance categories that could be chosen and in what direction the difference lay. To counter the inflation of the type-1 error, for each of three comparisons a significance level of $.02$ was used. The test values of each comparison are shown in Table 4.5.

It can be concluded that when the stance DP was depicted, the percentage of adjectives chosen that belong to the stance category DP did not significantly differ from the categories DH and SP, but the stance category SH had a higher percentage of chosen adjectives than DP. When the stance DH was depicted, the percentage of adjectives chosen that belong to category DH was significantly larger than the percentages of categories DP and SP. However, it did not differ from SH. When the stance SP was depicted the percentage of adjectives chosen that belong to category SP was significantly larger than the percentages of categories DP and DH, but significantly

Table 4.5: Testvalues are shown for several Wilcoxon Rank Sum tests in order to see which stances differ in percentages annotated adjectives given a certain acted stance.

		Chosen stance							
		DP		DH		SP		SH	
Act. st.	Cho. st.	Z	p	Z	p	Z	p	Z	p
DP	DP	x	x	-1.834	0.067	-0.762	0.446	-2.617	0.009
DH	DH	-4.136	$< .001$	x	x	-11.424	$< .001$	-0.975	0.330
SP	SP	-5.742	$< .001$	-5.148	$< .001$	x	x	-7.363	$< .001$
SH	SH	-14.369	$< .001$	-12.394	$< .001$	-12.375	$< .001$	x	x

smaller than category SH. Finally, when the stance SH was depicted, the percentage of adjectives chosen that belong to category SH was significantly larger than all the other categories.

In short, it was expected that the stance that was depicted should also have had the highest percentage of chosen adjectives. This was only the case for stance category SH. Actually, SH had the highest percentage (or shared the highest percentage of chosen adjectives) independently of the acted stance. If SH is not taken into account our expectations are met for the stances SP and DH and partially met for DP (which has a percentage of chosen adjectives that is equal to that of SP and DH). In the next section we will investigate this over representation of SH adjectives.

4.5.2 Distribution of Adjectives

Which adjectives did subjects select for the different categories of fragments? Table 4.6 shows how many times subjects used each of the 32 adjectives for fragments in the four categories. What does the table show? Compare SH adjectives ID $= 17$ 'defiant' and ID $= 24$ 'artificial'. Both were used frequently. But where artificial was used for all four stances in about the same number of judgements, defiant was used far more for fragments that expressed a DH stance. We also saw that 'helping' and 'loyal' fitted the DP stance better than 'leading', 'powerful', and 'stubborn', which were used more often to describe DH fragments. Adjectives 'gentle' and 'humble' fitted the SP stance best. Adjective 'reserved' fitted the SH stance best. Finally, adjectives 'offensive', 'impulsive' and 'cheeky' fitted the DH stance best.

Table 4.7 shows for each of the fragment categories the order of adjectives used in the judgements. The left-most column shows that over all fragments the adjectives 'doubting', 'reserved', 'defiant', 'artificial' and 'arrogant' were most frequently used. Remarkably all of them belong to the SH segment in Leary's Circumplex (adjective IDs $17 - 24$).

If we look at the SH column of Table 4.7 we see that all but one (namely 'depending') of the 8 SH adjectives that subjects could choose are in the top 8 of the list of most frequently selected adjectives for judgements of SH fragments. The DH column shows that out of the top 5 only 2 adjectives 'offensive' and 'cheeky' belong to the DH category. In the DP column none of the top 5 adjectives belong to the DP category.

Table 4.6: The adjectives (translated to English) ordered per stance and the counts of how many times subjects assigned each adjective to the fragments in each of the four categories.

Stance	ID.	Adjective	ALL	DP	SP	SH	DH
	1	leading	78	25	10	8	35
	2	powerful	113	24	7	11	71
	3	stubborn	106	20	11	26	49
DP	4	helping	47	28	9	5	5
	5	spontaneous	68	28	12	5	23
	6	lively	57	19	9	4	25
	7	loyal	51	26	13	6	6
	8	humane	41	15	15	7	4
	9	unprejudiced	34	11	8	11	4
	10	tolerant	30	14	4	9	3
	11	gentle	59	13	32	12	2
SP	12	cooperative	108	52	28	15	13
	13	humble	102	22	51	24	5
	14	discrete	79	15	27	32	5
	15	dead serious	83	19	23	13	28
	16	dependent	66	22	26	13	5
	17	defiant	198	24	19	54	101
	18	irreverent	120	21	19	45	35
	19	depending	73	10	32	28	3
SH	20	withdrawn	235	38	78	97	22
	21	arrogant	141	34	17	46	44
	22	doubting	236	50	96	71	19
	23	reserved	141	18	49	63	11
	24	artificial	166	39	38	50	39
	25	cynical	122	37	17	40	28
	26	compete	38	7	3	8	20
	27	bold	52	18	7	8	19
DH	28	offensive	102	12	6	4	80
	29	biased	81	20	10	22	29
	30	impulsive	75	16	9	6	44
	31	cheeky	115	20	11	25	59
	32	suspicious	83	15	28	26	14

The highest DP adjective is 'spontaneous' at rank 7. The 4 adjectives ranked $7-10$ all belong to this category, but 5 SH adjectives were more frequently used than these 4 DP adjectives. This is remarkable because these two categories are related to two segments of Leary's Circumplex that are diagonally opposite to each other. The SP column has one adjective in the top 5 that belongs to the SP category of adjectives (13. 'humble'). Also here 3 of the top 5 adjectives are SH adjectives.

Why did subjects find that the SH adjectives most appropriately described the stance taken by the actor? Is it because of the fact that they were 'artificial' behaviours was more apparent than the stance that was intended to be acted out? And was

Table 4.7: The adjectives (translated to English) ordered according to frequency of use (descending order) in the 672 judgements for all of the fragments and for the fragments from each of the four categories. Note that the numbers refer to the *identifing number* of the adjective, not to the count of occurences: adjectives for DP are identified with nrs 1-8, SP 9-16, SH 17-24, and DH 25-32.

ID	Adj(ALL)	ID	Adj(DP)	ID	Adj(SP)	ID	Adj(SH)	ID	Adj(DH)
22	doubting	12	cooperative	22	doubting	20	reserved	17	defiant
20	reserved	22	doubting	20	reserved	22	doubting	28	offensive
17	defiant	24	artificial	13	humble	23	withdrawn	2	powerful
24	artificial	20	reserved	23	withdrawn	17	defiant	31	cheeky
21	arrogant	25	cynical	24	artificial	24	artificial	3	stubborn
23	withdrawn	21	arrogant	11	gentle	21	arrogant	21	arrogant
25	cynical	5	spontaneous	19	depending	18	irreverent	30	impulsive
18	irreverent	4	helping	12	cooperative	25	cynical	24	artificial
31	cheeky	7	loyal	32	suspicious	14	discrete	18	irreverent
2	powerful	1	leading	14	discrete	19	depending	1	leading
12	cooperative	17	defiant	16	dependent	32	suspicious	29	biased
3	stubborn	2	powerful	15	dead serious	3	stubborn	25	cynical
13	humble	13	humble	17	defiant	31	cheeky	15	dead serious
28	offensive	16	dependent	18	irreverent	13	humble	6	lively
15	dead serious	18	irreverent	25	cynical	29	biased	5	spontaneous
32	suspicious	31	cheeky	21	arrogant	12	cooperative	20	reserved
29	biased	3	stubborn	8	humane	16	dependent	26	competitive
14	discrete	29	biased	7	loyal	15	dead serious	22	doubting
1	leading	15	dead serious	5	spontaneous	11	gentle	27	bold
30	impulsive	6	lively	31	cheeky	9	unprejudiced	32	suspicious
19	depending	23	withdrawn	3	stubborn	2	powerful	12	cooperative
5	spontaneous	27	bold	1	leading	10	tolerant	23	withdrawn
16	dependent	30	impulsive	29	biased	1	leading	7	loyal
11	gentle	32	suspicious	4	helping	27	bold	14	discrete
6	lively	14	discrete	6	lively	26	competitive	13	humble
27	bold	8	humane	30	impulsive	8	humane	16	dependent
7	loyal	10	tolerant	9	unprejudiced	7	loyal	4	helping
4	helping	11	gentle	2	powerful	30	impulsive	9	unprejudiced
8	humane	28	offensive	27	bold	5	spontaneous	8	humane
26	competitive	9	unprejudiced	28	offensive	4	helping	19	depending
9	unprejudiced	19	depending	10	tolerant	6	lively	10	tolerant
10	tolerant	26	competitive	26	competitive	28	offensive	11	gentle

'doubting' the stance that an actor expresses when he/she intends to act a certain stance but doubting how to express this? This raises the question of whether using acted stance fragments is a good idea for studying whether people agree in describing the stance that someone takes. On the other hand it could also indicate that people were biased towards interpreting stances as SH, and that the accompanying adjectives were their default opinion.

4.5.3 Individual Judgements

In total 84 subjects each judged 8 fragments by selecting at least 4 adjectives—from a list of 32—that they found most appropriately described the stance acted by the actor shown in the fragment. So in total we have $84 * 8 = 672$ judgements. Subjects were asked to choose at least 4 adjectives, no maximum was set. Per fragment subjects used 4.6 adjectives in the mean with a maximum of 10 adjectives. In 434 judgements subjects selected four adjectives, in 134 judgements 5 adjectives were selected, in 62 judgements 6 adjectives, in 24 judgements 7 adjectives, in 10 judgements 8 adjectives, in 6 judgements 9 adjectives, and in only 2 judgements 10 adjectives were selected. Since the adjectives belong to one of 4 categories of Leary's Circumplex, it is interesting to see how often there was a match between the category of the adjective chosen and the category of the stance acted out in the fragment. A judgement of a fragment by a subject is called:

- Perfect, when *all* the adjectives that a subject has chosen as describing that fragment belong to the same class as the stance that was intended by the actor in the fragment.
- Correct, when there is a unique category with a maximum number of adjectives selected (a unique majority category) and this category is the same as the intended stance of the fragment.
- Semi-correct, when the category that has the maximum number of adjectives chosen is the same as the intended stance of the fragment.
- Wrong, if it is not semi-correct, correct, or perfect.

Note that when a judgement is perfect it is also correct and when it is correct it is also semi-correct. Thus, a judgement is either semi-correct or wrong and their sum is the total number of fragments of a class. A judgement that has for example 4 adjectives 2 of which are of the intended stance and 2 are of another stance category is semi-correct. It is not correct since it has no unique majority category.

4.5.3.1 Results for all Fragments

Table 4.8 shows for each of the categories how many times the judgements were perfect, correct, semi-correct, or wrong. The total number of judgements is 672. There are small differences in the numbers of fragments in each of the four categories. From the total of 672 judgements 162 judgements concern DP fragments, 178 concern DH fragments, 157 concern DP, and 175 SH fragments. Table 4.9 shows how many adjectives subjects assigned to the subsets of fragments of the four different stance categories. For example in total 454 adjectives of the category SH were assigned to the fragments of category SH, but 234 of these SH adjectives were assigned to the fragments of stance category DP. It is clear that by far most of the adjectives selected by the subjects belong to the category SH. This explains the outstanding number of perfect judgements made for the SH fragments (see Table 4.8).

Table 4.10 shows the confusion matrix. It shows for each of the stances (rows) how often fragments of that stance were assigned the four classes if we take the

Table 4.8: The number of times subjects assigned the 'correct' stance to the fragments in each of the four categories. For explanation of what 'correct' means see the main text.

	Judgements				
CAT	PERF	CORR	SEMICOR	Wrong	Total
DP	3	28	53	109	162
SP	1	27	47	110	157
SH	22	113	138	37	175
DH	0	52	84	94	178

Table 4.9: The number of times subjects assigned the stance adjectives indicating the different stance categories to the fragments in each of the four categories.

	Chosen Stance			
CAT	DP-A	SP-A	SH-A	DH-A
DP	185	168	234	145
SP	86	199	348	91
SH	72	129	454	139
DH	218	65	274	293

stance category with the maximum number of adjectives as the stance assigned. In cases where there was no unique stance category with a majority then the decision is X (undecided). From the numbers in Table 4.10 we computed the precision, recall and F-values (Table 4.11). The SH and DH (both hostility) categories have clearly higher F-measures than the two categories DP and SP (both positivity). The highest precision was obtained for class DH.

4.5.3.2 Clustering Judgements

To further investigate the structure of the ratings we performed a k-means clustering (with $k = 4$ as there were four stances) of the ratings for every adjective (32 adjectives, so a 32-dimensional binary vector). Every rating was a case (so there were 672 cases) and each fell in one of the four clusters (CL1-CL4). For every case we know the intended stance of the actor, so we can see how often each acted stance occurred in each cluster (Table 4.12). We gave the cluster the name of the stance that occurred most in that cluster. We compared the F-values of the clustering (Table 4.13) and the F-values of the assignments by majority vote (Table 4.11): this showed us (unsurprisingly) that by clustering the F-values went up. This means that the ratings of the judges seem to be structured differently by the clustering compared to the stance categories. With this method we can only see to what extent the judges chose the same adjectives, we cannot investigate whether the different judges chose adjectives

Table 4.10: The number of times subjects assigned stances to the fragments in each of the four categories or if the chosen stance was undecided.

	Chosen Stance				
CAT	DP-C	SP-C	SH-C	DH-C	X-C
DP	28	**31**	43	20	40
SP	13	27	77	7	33
SH	6	15	**113**	14	27
DH	**36**	6	49	**52**	35

Table 4.11: The accuracy, precision, recall and F-values for each of the four stance categories. These figures are based on the figures in the confusion Table 4.10.

CAT	Acc	Precision	Recall	F
DP	0.72	0.34	0.17	0.24
SP	0.73	0.34	0.17	0.24
SH	0.66	0.40	0.65	0.52
DH	0.75	0.56	0.29	0.38

from the same stance category. Judges might have disagreed on the adjectives that they chose, but those adjectives might have (largely) belonged to the same stance category.

Table 4.12: Counts how many acted stances occur in the clusters. The majority of occurrences is the name for that cluster.

	Counts of Ratings in each Cluster				
CAT	CL3-DP	CL4-SP	CL2-SH	CL1-DH	Total
A-DP	**64**	30	38	30	162
A-SP	34	**88**	23	12	157
A-SH	20	81	**59**	15	175
A-DH	10	10	43	**115**	178
Total	128	209	163	172	672

To test this, we compared the clustering of the adjectives (32-dimensions) to a clustering of a 4-dimensional vector that represented the stance grouping of the adjectives. For every case, we counted how many adjectives from the group of adjectives of the stance were chosen. This resulted in a 4D vector (each stance) with values that ranged from 0 (no adjective from that stance was chosen) to 8 (all adjectives from that stance were chosen). We performed a k-means ($k = 4$) on this 4-dimensional

Table 4.13: The precision, recall and F-values of the 4-means clustering based on table 4.12.

CAT	Precision	Recall	F
DP	0,50	0,40	0,44
SP	0,42	0,56	0,48
SH	0,36	0,34	0,35
DH	0,67	0,65	0,66

representation of the data. Again, for every case we know the intended stance of the actor and can see how often each acted stance occurred in each cluster (Table 4.14). We gave the cluster the name of the stance that occurred most in that cluster. We compared the F-values of the two clusterings (tables 4.13 and 4.15): the F-values for the 4D clustering were lower for stances DP, SP and DH, and higher for SH. So by clustering over the 4D stances the performance dropped. This means that the choices of the judges did not fall into the same stance categories, but were distributed over different stance categories.

Table 4.14: Counts how many acted stances occur in the clusters on the 4(stance)-dimensional data. The majority of occurrences is the name for that cluster.

	Counts of Ratings in each Cluster				
CAT	CL3-DP	CL4-SP	CL2-SH	CL1-DH	Total
A-DP	**45**	42	27	48	162
A-SP	18	**53**	53	33	157
A-SH	9	38	**85**	43	175
A-DH	44	9	32	**93**	178
Total	116	142	197	217	672

Table 4.15: The precision, recall and F-values of the clustering based the 4(stance)-dimension data, based on table 4.14.

CAT	Precision	Recall	F
DP	0,39	0,28	0,32
SP	0,37	0,34	0,35
SH	0,43	0,49	0,46
DH	0,43	0,52	0,47

4.5.3.3 Results for Sound and Mute Fragments

There were 336 S-fragment (sound/with audio) judgements, the same as the number of M-fragment (mute/no audio) judgements. Tables 4.16 and 4.18 show the scores and the assignment of adjectives for the part of the corpus with sound. Tables 4.17

and 4.19 show the scores and the assignment of adjectives for the part of the corpus without sound.

In Figure 4.5 the judgements with sound, muted and total are visualised in a bar graph. The total value has been divided by 2 which represents the judgements if sound and mute were to be fully equally distributed. From this graph we can compare the judgements of the S- and the M-fragments and we see that there are no significant differences between the fragments with and without audio. The percentages of wrong judgements is the same for the fragments with audio and without audio. This holds for all 4 stances.

Table 4.16: The number of times subjects assigned the 'correct' stance to the S-fragments in each of the four categories. For explanation of what 'correct' means see the main text.

	Judgements Sound Fragm.				
CAT	PERF	CORR	SEMICOR	Wrong	Total
DP	1	15	33	57	90
SP	0	13	23	51	74
SH	11	60	68	17	85
DH	0	25	43	44	87

Table 4.17: The number of times subjects assigned the 'correct' stance to the M-fragments in each of the four categories. For explanation of what 'correct' means see the main text.

	Judgements Mute Fragm.				
CAT	PERF	CORR	SEMICOR	Wrong	Total
DP	2	13	20	52	72
SP	1	14	24	59	83
SH	11	53	70	20	90
DH	0	27	41	50	91

4.5.3.4 Results for the Theory Actors

We had two different conditions in which actors were asked to perform the four stances. Four of the eight actors were recorded in the 'Theory condition' (T-condition). The other four in the scenario or 'Role play condition' (R-condition).

Are there differences between these two groups of actors if we look at how many judgements were correct? Or, in other words did subjects recognize the intended stances better when this stance was acted in the T-condition than when the stance was acted in the R-condition? Half of the judgements (336) involve actors in the T-condition, the other half in the R-condition.

Table 4.18: The number of times subjects assigned stances to the S-fragments in each of the four categories.

	Chosen Stance Sound Fragm.				
CAT	DP-C	SP-C	SH-C	DH-C	X-C
DP	15	16	19	10	30
SP	4	13	38	3	16
SH	2	6	60	7	10
DH	13	1	28	25	20

Table 4.19: The number of times subjects assigned stances to the M-fragments in each of the four categories.

	Chosen Stance Mute Fragm.				
CAT	DP-C	SP-C	SH-C	DH-C	X-C
DP	13	15	24	10	10
SP	9	14	39	4	17
SH	4	9	53	7	17
DH	23	5	21	27	15

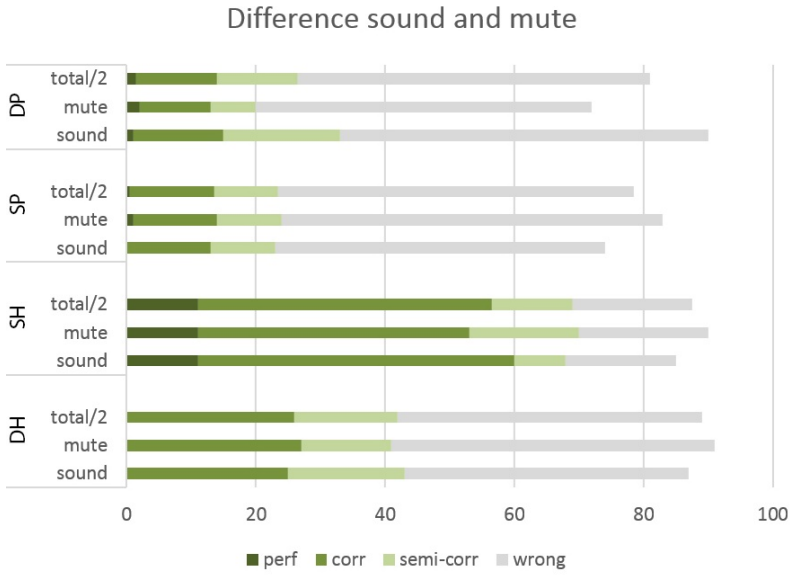

Figure 4.5: Difference between mute and sound.

The Tables 4.20 and 4.22 show the results for the T-actors. The Tables 4.21 and 4.23 show the results for the R-actors. In Figure 4.5 the judgements with sound, muted, and total are visualised in a bar graph. The total value is divided by 2 which represents the judgements if sound and mute were to be fully equally distributed. If we compare the figures in Tables 4.20 and 4.21 we see that of the 22 perfect judgements involving an acted SH stance, 19 were performed in the Theory condition. Only 3 in the Role play condition. In Figure 4.6 the judgements for the Theory condition, Role play condition and total are visualised in a bar graph. The total value is divided by 2 which represents the judgements if Theory and Role play were to be fully equally distributed. As can be seen here it appears that overall the Theory conditions induces acted stances that are better recognized than those in the Role play condition.

Table 4.20: The counts how many times subjects assigned the 'correct' stance to the T-fragments in each of the four categories. For explanation of what 'correct' means see the main text.

	Judgements Theory Fragm.				
CAT	PERF	CORR	SEMICOR	Wrong	Total
DP	2	20	32	43	75
SP	0	11	19	60	79
SH	19	74	83	10	93
DH	0	31	43	46	89

Table 4.21: The counts how many times subjects assigned the 'correct' stance to the R-fragments in each of the four categories. For explanation of what 'correct' means see the main text.

	Judgements Role Play Fragm.				
CAT	PERF	CORR	SEMICOR	Wrong	Total
DP	1	8	21	66	87
SP	1	16	28	50	78
SH	3	39	55	27	82
DH	0	21	41	48	89

4.5.3.5 Preliminary Conclusion Individual Judgements

The results presented in this section show that there is a clear correlation between the stance of the fragments and the adjectives chosen by the total of all subjects in their judgements of these fragments. This is clear from the distribution, Table 4.6. The observed frequencies on the main diagonals of these tables are always considerably larger than their expected values. The $\chi^2(df = 9)$ values are respectively 383, 254 and

Table 4.22: The counts how many times subjects assigned stances to the T-fragments in each of the four categories.

	Chosen Stance Theory Fragm.				
CAT	DP-C	SP-C	SH-C	DH-C	X-C
DP	20	9	17	8	21
SP	10	11	39	2	17
SH	1	3	74	6	9
DH	9	1	35	31	13

Table 4.23: The counts how many times subjects assigned stances to the R-fragments in each of the four categories.

	Chosen Stance Role Play Fragm.				
CAT	DP-C	SP-C	SH-C	DH-C	X-C
DP	8	22	26	12	19
SP	3	16	38	5	16
SH	5	12	39	8	18
DH	27	5	14	21	22

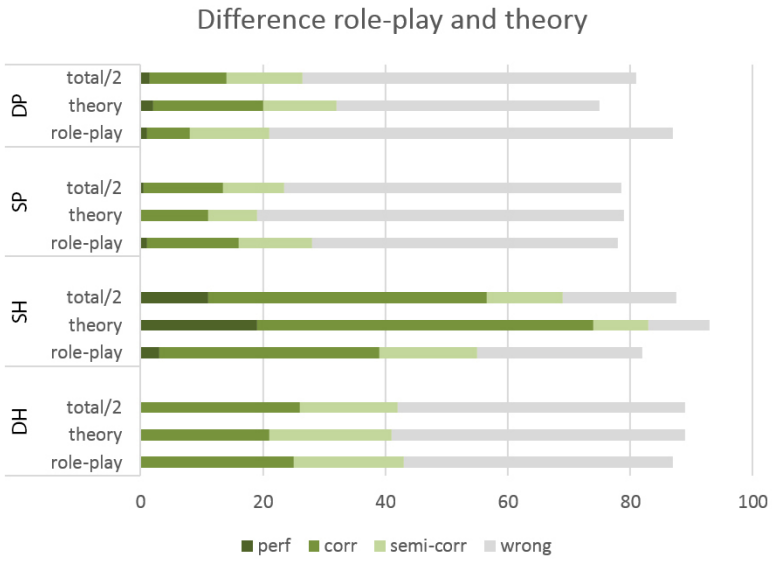

Figure 4.6: Difference between theory and role-play groups.

171 all with $p << 0.0001$. Of course this is as it should be. Ideally all values should be on the main diagonal. On the other hand there were many judgements in which subjects chose adjectives that belong to a different category than the stance category that was intended by the actor. Up to now we have only analysed which adjectives the group of all subjects used to describe the fragments of the various stance classes. It is quite possible that in cases where subjects did not recognize the stance as it was intended, they were at least in agreement with each other.

4.5.4 Inter-annotator Agreement

Did subjects agree on the use of adjectives for the different fragments? If subjects did not agree in their accounts of the stance taken by the actors in the fragments they judged then it is difficult to say what the stance was that the actor took. The meaning of 'content' in the discipline of content analysis is not always clear [111]. Here content is something of the interaction between what is presented, in this case the behaviour shown in the video fragments and the subject who had the task to describe the stance taken by the actor. There is no 'golden truth'. If most of the subjects who judged a certain fragment judged this as a *submissive* stance then this is something we have to take as 'content' even if the stance that was intended by the actor was more of the type *opposing* than *submissive*. We analysed the judgements for inter-rater agreement. Our 'coding task' had the following properties:

- There was a large number of annotators (84).
- Not all annotators annotated all fragments. There was a total of 64 fragments: 8 actors, acted each 4 stances, and we have all recordings with and without audio, makes $2 * 8 * 4 = 64$ fragments. Each of the annotators labelled 8 fragments. The fragments were assigned to the annotators in a random way, but ensuring that not all fragments belonged to the same category.
- The label set used in the annotation task was large. A subject could label a fragment with any subset of the set of 32 adjectives, with the restriction that it contained at least four adjectives.

Because of these properties we used Krippendorff's α agreement method for computing a reliability measure [111, p222]. For a thorough discussion about the various methods and measures for inter-rater agreement see [5]. We applied the method for many observers, many nominal categories, and for missing values [111, p232].

Since the annotators assigned a set of adjectives to each fragment, category labels in this annotation task consist of sets of adjectives. However, since the number of potential labels was quite large (potentially as many as there were subsets -with at least 4 elements- of a set of 32 elements) and many of them had not been used, we decided not to use sets of adjectives, but sets of stance categories. Note that annotators did not know the stance categories of the adjectives; the adjectives were given in a random order. Theoretically we now had $2^4 - 1 = 15$ different labels. They correspond to the non-empty subsets of the four stance categories.

Krippendorff's α allows us to plug in a distance metric on the label set. We used a difference metrics based on the similarity measure on sets known as *Dice coefficient*,

Table 4.24: The α values computed for fragments with and without sound and for the role play and theory fragments using Krippendorff's method with Dice metrics for distances between values.

	α - audio		α - condition	
ALL	Mute	Sound	Role Play	Theory
0.22	0.21	0.23	0.15	0.27

Table 4.25: The α values computed for the fragments of each acted stance and for all fragments excluding the fragments of a specific acted stance using Krippendorff's method with Dice metrics for distances between values.

α - stance categories							
DP	-DP	SP	-SP	SH	-SH	DH	-DH
0.12	0.23	0.08	0.24	0.03	0.21	0.22	0.15

see [5, 57, 126]. Suppose two annotators assigned sets C_1 and C_2 each containing 4 adjectives to a certain fragment. In each of the two sets 2 adjectives belonged to stance category X, the other two belong to Y. This means that both annotators were in a sense uncertain about the stance expressed. But the metric does not penalize this as disagreement. To give another example: the distance between the sets $\{DP, DH, SP\}$ and $\{DH, SP\}$ is 0.2. The results of the inter-rater agreement analysis are shown in Table 4.24. It shows the α values for the whole class of fragments and for the class of S-fragments (with audio) and the class of M-fragments with muted audio. These values are low. There was slightly more agreement on the fragments with audio than there was on the fragments without audio. Clearly, the Theory play judgements had a higher inter-rater agreement. We also computed α for parts of the corpus containing only fragments of a certain intended stance, see Table 4.25. This table also shows the α values for the corpus without the parts containing fragments of a specific stance. The exceptional values for the DT fragments are remarkable. Remember that this was the class that also had the highest precision value. DT stance behaviour is easier to recognize (and perform!) than the other types of stances.

4.5.5 Were some actors better than others?

We saw that stances acted in the Theory condition were better recognized and had a higher inter-rater agreement than the stances acted in the Role play condition. In each condition 4 actors performed the stances. Now we will look at individual actors. Were some actors better than others in the sense that the stances they performed were more easily recognized by the subjects? To answer this question we computed a score for each of the actors. For each of the actors we looked at the judgements

Table 4.26: The scores and α reliability values for each of the 8 actors

Actors in Theory-Condition							
T01		T02		T03		T04	
score	α	score	α	score	α	score	α
72	0.16	90	0.22	58	0.03	114	0.38
Actors in Role Play-Condition							
R01		R02		R03		R04	
score	α	score	α	score	α	score	α
61	0.21	68	0.24	62	0.04	43	0.01

in which the actor acted. If the judgement was perfect we added 3 points to the score, if it was correct we added 2 points to the score, if it was semi-correct we added 1 point to the score. Since all actors were involved in the same number (84) of judgements we did not have to normalize these scores. The resulting scores are in Table 4.26. The T-actors are the actors that acted in the Theory condition, the R-actors are those that acted in the Role play condition. Actor T04 scored significantly higher than the mean score and actor R04 scored lower than the mean. What is the impact of these two actors on the $alpha$ values? If we remove all judgements with R04 α slightly raises from 0.22 to 0.25. If we remove T04 α becomes 0.19. If we only take the fragments with actor T04 α raises to 0.38. Our analysis confirms that some actors were better than others and that good acting had a significant impact on inter-annotator agreement.

Since fragments were assigned randomly to subjects there is a chance that the positive and negative scores for T04 and R04 were due to the subjects, not to the actors. In order to cancel this out we looked at those judgements by subjects that both annotated the same stance by the same actors (for actors T04 and R04). Do these judgements differ in quality if we vary the subject or if we vary the actor?

The data contains 8 subjects that annotated both actors R04 and T04 acting stance SH, 3 subjects that annotated both actors acting stance SP, 7 for stance DT and 6 for stance DP. In total 24 different subjects annotated the two actors acting the same stance. For each of these 48 judgements we computed the scores and we analysed the results. The scores for R04 has a mean of 0.42 (SD 0.88) and for T04: mean is 1.42 (SD 0.93). A paired t-test comparing scores for R04 and T04 on the 24 pairs of judgements of the same stances by the same subjects gives: $t(23) = 4.796$ ($p << 0.0001$). In all but one case the judgement of a subject has a higher score with actor R04 than with actor T04. In all other cases T04 scores equal (9 times) or higher (14) than R04. This gives sufficient evidence to rule out that the higher scores for T04 compared to those for R04 are due to the judges assigned to them. Table 4.26 also contains the α reliability values for the 8 parts of the corpus divided per actor. Figure 4.7 shows the relation between scores and α values. It shows that the ratio between scores

and α values varies considerably. For actors $R01$ and $R02$ they are much higher than for R03. The Spearman correlation between α and score equals 0.833 (significance $p < 0.05$, 2-tailed). Overall, there is a reasonable correlation between the inter-annotator agreements and the validity. But as the correlation graph shows, for some actors (e.g. R03) a higher score (validity: agreement between judgement and intended stance) goes with a low inter-annotator agreement and for others (e.g. R01, R02) a lower score (validity) goes with a higher inter-annotator agreement meaning relatively more annotators agree on the stance they see but it is not the stance as it was intended. We see that an actor being good has two different senses: he performs the stance that he was asked to act, or he performs a stance that is recognized by a majority of observers. We see that in the mean the T-actors have higher scores as well as higher inter-annotator agreement than the R-actors.

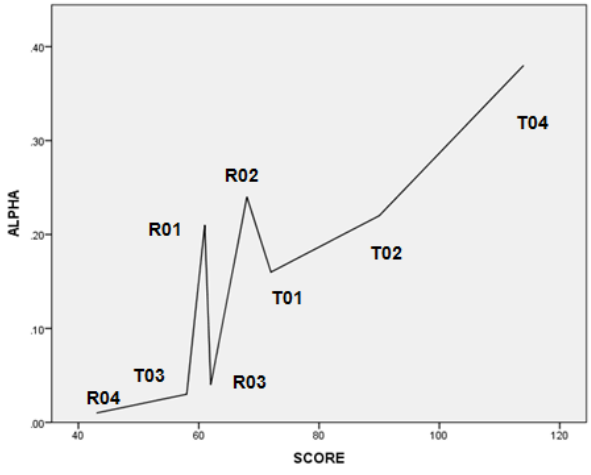

Figure 4.7: Judgement scores and inter-annotator agreement for each of the 8 actors.

4.5.6 Best Fragments

This research was conducted to contribute to a project with a wider perspective. Namely to bring forth a conversational agent that can be used to train interrogation skills for police men and women. The part this research will take in the bigger project is to try to describe certain postures which could be depicted by a conversational agent. If these different postures are valid, they can be used to provoke certain reactions according to Leary's interpersonal stance relations [116]. As described before, the relation between depicted stance and perceived stance seems very weak. This is why it could be difficult to clearly define a typical and valid posture that depicts a certain stance. Nevertheless we will try to qualitatively describe the best fragment of each depicted stance. In order to determine which fragments are the best, all 4 stances of all 8 actors, which represents all fragments, were judged and plotted. The

judgement of these fragments was done by inter-rater agreement and the score system as used before. This plot is shown in Figure 4.9. For practical reasons the actors are numbered consecutively where actor number 1 till 4 represent T01 till T04 and 5 till 8 represent R01 till R04. As can be seen in this plot alpha reliability values are very low. This is probably mainly caused by the small number of respondents on each separate fragment. As described by [111] these values are far from relevant and therefore will not be used in the judgement of the fragments. When only taking the judgement scores into account the selected fragments can be seen in Table 4.27. Figure 4.8 shows stills for the best actors for the four stances. There is a similarity observable between the stills of different actors for the same stance. For example, raised eye-brows for DP, downward gaze for SP, folded arms for SH, and raised arms for DH.

Figure 4.8: The best actors for each of the four stances. Observe the apparent similarity across different actors within the same stances.

Table 4.27: Selected fragments.

Stance	English description	Actor nr.
DP	Dominant-Positive	4
SP	Submissive-Positive	5
SH	Submissive-Hostile	2
DH	Dominant-Hostile	4

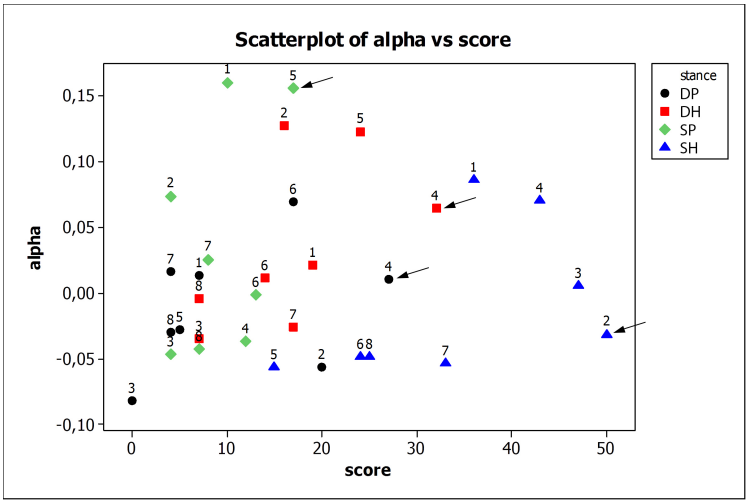

Figure 4.9: Judgement scores and inter-rater agreement for each of the 8 fragments.

4.6 Effects of Actor Proficiency

Some of the actors in the previous study mentioned they found the task of acting out an interpersonal stance difficult and unnatural. This might have influenced the judgements in such a way that the SH-adjectives were chosen as the most appropriate for most acted stances. Their behaviour felt 'artificial' which was one of the adjectives describing the SH stance. The fact that the behaviour was 'artificial' might have been more apparent than the stance that was intended. The question is whether the (lack of) expertise of our actors might have obfuscated the intended stance with behaviour that is interpreted as SH. To address this issue, we repeated the study using clips taken from TV-shows that feature police interrogations. The idea here was that the professional actors in these clips were better at acting. In addition, we asked raters explicitly about the spontaneity of the behaviour of the actor in the fragment.

4.6.1 Professional Actor Fragments

We selected fragments from TV-series. The interpersonal stance of the suspect in each fragment was determined from the content of the entire episode. This could mean anything from the content of what they said to explicit comments made by the characters in the episode. Observing the (non verbal) behaviour of the actors, we kept in mind the typical stance behaviours from the literature, see Table 4.3. We categorised the stance in these fragments using our best judgement, however we have seen from previous work on stance judgements that this subjective task often has a low inter-rater agreement (e.g. [141]). In other words, we and the participants in our study might not agree on the stance that is portrayed in a fragment. This introduces uncertainty about which stances the fragments that we selected would actually depict

according to a majority vote of multiple observers. We suggest that a majority vote on stance would be closer to the stance that is portrayed than an expert opinion. We assessed which stance was actually depicted in the fragments, see Section 4.6.3.

To ensure that the fragments were similar and comparable to the fragments used in the experiment described earlier, the fragments had to meet three criteria: the suspect is the only one in the fragment, the suspect is being interrogated (seated in a room), and the length of the fragment was similar to the acted stances in the previous study ($3 - 10$ seconds). Figure 4.10 shows stills of some of the video fragments from TV-series used in the study. The actors show different stances (see caption). We selected four fragments for every stance: three from professional actors and the best recognized fragment from the previous study with our amateur actors. These fragments allow us to compare both data-sets.

Figure 4.10: Stills from the fragments with the professional actors. a. Dominant-hostile expression: tilt head up with a gaze down toward the interrogators; b. Dominant-hostile posture: asymmetrical, space-filling and distant posture; c. Submissive posture: shrinking posture with bent spine; d. Hostile expression: expression of contempt; e. Dominant expression: expressive face with extreme brow raising; f. Submissive expression and posture: expression of sadness and self-touch; g. Friendly expression: smile; h. Hostile expression: cross-eye gaze with eyelid raising.

4.6.2 Method

Participants observed each fragment and rated them on the two interpersonal stance dimensions, dominance and affiliation, and also on spontaneity. Instead of having participants label the fragments with 32 adjectives we opted for rating on the stance dimensions to mitigate concerns about the ambiguity that using adjectives might have introduced, see Section 4.5.1. To assess the quality of acting we asked how spontaneous the behaviour in the fragment appeared. We used 5-point Likert scales. The labels on the scales were: very dominant (5) - very submissive (1) for dominance,

very friendly (5) - very hostile (1) for affiliation, and very spontaneous (5) - very acted (1) for spontaneity.

4.6.3 Fragments' Stance

In total 65 participants (aged: $mean = 25.4$, $min = 14$, $max = 50$, and $SD = 6.2$. Gender: 31 female.) judged the 16 fragments on 5-point Likert scales on the three dimensions: dominance, affiliation, and spontaneity. If the mean of the judgements of all participants was above or below the midpoint (which is 3) we classified the fragment in the respective category. For example, if for a fragment the mean on the dominance scale was above 3 it was rated as dominant, whereas the mean was below 3 the fragment was rated as submissive. This analysis can show us the stance that was depicted in each fragment. Table 4.28 shows our prediction of stance versus the result of this categorization of the fragments based on the judgements of the participants. We concluded that we had 3 fragments that depict a DP stance, 6 DH, 4 SP, and 3 SH. Note that this includes the fragments from our amateur actors, see section 4.6.4.

Table 4.28: Our prediction of stance versus the categorization of the fragments based on the judgements of the participants.

		Outcome			
		DP	DH	SP	SH
Predicted	DP	3	0	1	0
	DH	0	4	0	0
	SP	0	1	3	0
	SH	0	1	0	3
	Total	3	6	4	3

4.6.4 Amateur Actors

We know the intended stance from the fragments with our amateur actors. Fragments 5, 9, 13, and 17 feature our amateur actors, see Table 4.29. Fragment 5 was acted with a DP stance and the mean rating of dominance was above 3 meaning it was rated as dominant, and the mean rating of affiliation was above 3 meaning it was rated as positive: a DP stance. Fragment 9 was acted with a DH stance and the mean rating of dominance was above 3 meaning it was rated as dominant, and the mean rating of affiliation was *below* 3 meaning it was rated as hostile: a DH stance. Fragment 13 was acted with an SP stance and the mean rating of dominance was below 3 meaning it was rated as submissive, and the mean rating of affiliation was above 3 meaning it was rated as positive: an SP stance. Fragment 17 was acted with an SH stance and the mean rating of dominance was below 3 meaning it was rated as submissive, and the mean rating of affiliation was below 3 meaning it was rated as hostile: an SH stance. The mean ratings match the stance that was intended by the amateur actors for all of these fragments.

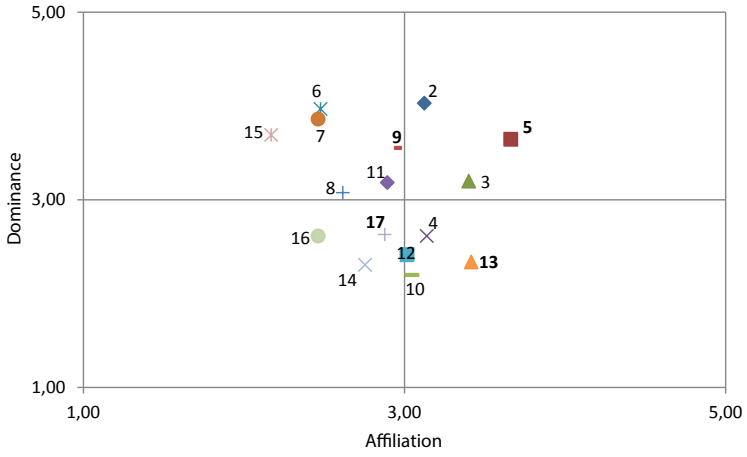

Figure 4.11: The location on the interpersonal circumplex of all the fragments based on the mean ratings on dominance (y-axis) and affiliation (x-axis). Note that fragment 1 was used to familiarize participants with the procedure and is excluded from analyses. Fragments 5, 9, 13, and 17 are the amateur actors and are displayed bold.

We were concerned that the expertise of our actors might have influenced the perceived stance. In the previous perception study we did not explicitly ask our participants to rate the quality of acting. In this experiment we did, see Table 4.29, and it showed that none of our amateur actors were rated with a mean over 3 points (the mid-point) for *spontaneity*. From the fragments with professional actors only 2 fragments were rated with a mean score below 3 and 10 were rated with a score of 3.03 or higher. This means that the behaviour of the amateur actors was rated as 'acted' and not as spontaneous, where the behaviour in most professional fragments was rated as spontaneous rather than acted. In the previous experiment we found that our amateur actors were described most with adjectives from the SH stance, see section 4.5.1. The results make us conclude that the expertise of our actors might have influenced the perceived stance. This conclusion is in line with earlier findings from [37] who suggest that professional actors may provide a more natural representation of interpersonal emotions avoiding exaggeration or caricature behaviours.

4.6.5 Spontaneity and Inter-rater Confusion

We can further investigate the relation between the expertise of the actor and the judgements on dominance and affiliation. For this we compare the ratings on spontaneity with the standard deviation on dominance and affiliation for all fragments. The correlation between these measures tells us something about the influence the spontaneity had on the clarity of the acted behaviour. The reasoning is that 'unclearly' acted behaviour will lead to a larger deviation as raters have to guess, they disagree more when the behaviour of an actor is inconsistent or conflicting.

Table 4.29: Mean ratings for the fragments for dominance, affiliation, and spontaneity. The fragments with our amateur actors (fragments 5, 9, 13, and 17) have the stance as intended by the actor displayed. Note that fragment 1 was the example fragment and it has not been included in the analyses.

Fragm.	Dom. Mean	Aff. Mean	Stance Result	Spont. Mean
2	4.03	3.12	DP	3.31
3	3.20	3.40	DP	3.06
4	2.62	3.14	SP	3.42
5 - DP	**3.65**	**3.66**	**DP**	**2.65**
6	3.97	2.48	DH	3.05
7	3.86	2.46	DH	3.03
8	3.08	2.62	DH	3.22
9 - DH	**3.55**	**2.94**	**DH**	**2.45**
10	2.20	3.05	SP	3.14
11	3.18	2.89	DH	2.97
12	2.42	3.02	SP	3.09
13 - SP	**2.34**	**3.42**	**SP**	**2.94**
14	2.31	2.75	SH	3.06
15	3.69	2.17	DH	3.14
16	2.62	2.46	SH	3.05
17 - SH	**2.63**	**2.88**	**SH**	**3.00**

Table 4.30: Spearman's rho (data is non-parametric) correlations for spontaneity and rater confusion measured by the deviations on affiliation and dominance.

		Dev. of Aff.	Dev. of Dom.
Spont.	Corr. Coef.	-.398	-.517*
	Sig. (2-tailed)	.127	.040
	N	16	16

Correlation and regression analyses were conducted to examine the relationship between the predictor spontaneity (quality of acting) and deviation of dominance and affiliation (confusion between raters). Figures 4.12 and 4.13 show the scatterplots for spontaneity and the deviation of affiliation and dominance. It seems that indeed there was a trend (albeit weak) that an increase in spontaneity decreased the deviations. Table 4.30 shows that spontaneity significantly predicted the confusion for dominance ($p < 0.05$). However, spontaneity was not a significant predictor for the deviation of affiliation ($p > 0.1$).

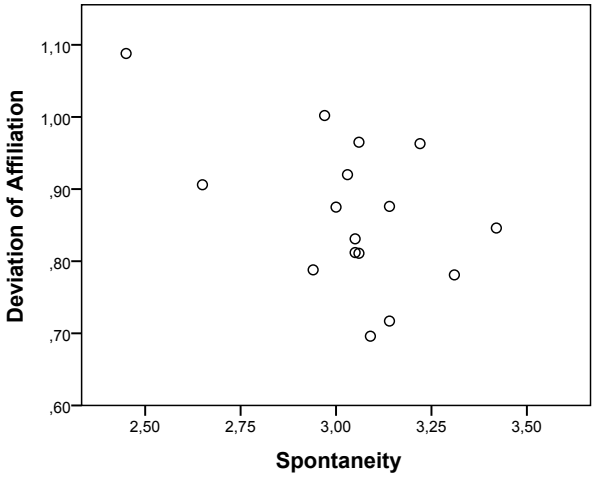

Figure 4.12: The scatter-plot for the spontaneity and the deviation of affiliation for all fragments.

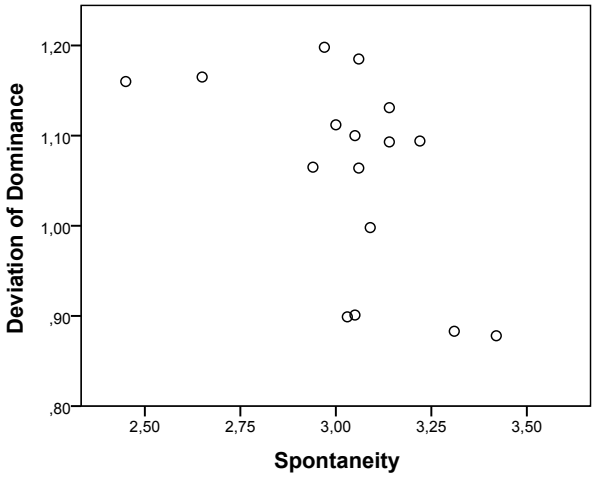

Figure 4.13: The scatter-plot for the spontaneity and the deviation of dominance for all fragments.

4.6.6 Clustering of Ratings

To further investigate the structure of the ratings we again performed a k-means clustering (with $k = 4$ as there were four stances) on the ratings on dominance, affiliation and spontaneity for each of the fragments. There were 17 clips rated by 65 judges resulting in 1105 cases that were rated on three scales (3-dimensions). Earlier in this section we determined the stance that was *perceived* in each clip. We counted how

Table 4.31: Ratio of counts of judgements in each cluster divided by total judgements for that stance (see main text for details).

	Cluster Judgement Fractions			
CAT	CL1-DP	CL3-SP	CL2-SH	CL4-DH
DP	**0,523**	0,036	0,185	0,256
SP	0,265	**0,269**	0,377	0,088
SH	0,215	0,236	**0,385**	0,164
DH	0,422	0,075	0,158	**0,345**
Total	1,426	0,616	1,104	0,854

Table 4.32: The precision, recall and F-values of the clustering, based on table 4.31.

CAT	Precision	Recall	F
DP	0,367	0,523	0,431
SP	0,437	0,269	0,333
SH	0,348	0,385	0,366
DH	0,404	0,345	0,372

many times each stance occurred in each cluster. However, we did not have an equal number of clips for each stance. Therefore we divided the count of how many times the stance occurred in each cluster by how many fragments of that stance were rated, giving the ratio of the counts in the cluster and the occurrences of the stance (Table 4.31). We gave the cluster the name of the stance that occurred (relatively) most in that cluster. The precision, recall and F-values for this analysis (Table 4.32) were similar to those for the amateur actor ratings. This means that the confusion of raters was similar for amateur and professional actors.

4.7 Conclusions

The background idea of the stance perception studies is that there are typical 'tiny behaviours' that humans make in a conversation that together make up how their interpersonal stance is perceived by the observer. The idea is quite common, see the references in Table 4.3. The analyses of our perception studies show once again that there is a complex relation between the isolated observable elements (posture, gesture, or facial expressions) on the one hand and the perceived stance on the other.

The results of the annotations show a clear correlation between the stances that were acted in the videos and the adjectives that were chosen in the judgements. However, there were many judgements in which subjects chose adjectives that belonged to another category than the stance category that was intended by the actor. We see that inter-annotator agreement is low ($\alpha = 0.22$). Other studies that looked at inter-annotator agreement in a stance annotation task using Leary's Circumplex have already shown that this is a difficult task (see, e.g., [199] and the previous chapter).

The stance that was seen most in the first perception study is Submissive-Hostile (SH) independent of the stance that was intended by the actor. This can first of all be because the interpersonal stances in the videos were acted and several actors commented that the task felt unnatural to them which could have influenced the naturalness of their acting. Secondly, this could indicate that raters have a default opinion about the clips they are judging. The raters were explained they were going to watch suspects in a police interrogation. Given this context raters might have been biased in their observations thinking that suspects would have acted hostile and submissive towards a police officer, since they were being charged of doing something wrong by a dominant public figure. This research therefore gives an indication that information about the setting in which communication takes place, can guide people in their observations. Whether the bias found towards SH really indicated a bias provoked by the setting or whether this was a resultant of using actors, needs to be further investigated. It should also be investigated whether different biases are found when different communicative settings are used. In this experiment audio did not add necessary information for annotating interpersonal stance. It has to be noted that some videos contained silent acting. The judgements of actors in the theory-condition differed from the judgements in the role play-condition. For most stances, fragments with actors from the theory-condition were better recognized. This is most obvious with the acted stance SH where 19 of 22 perfect judgements are in the theory-fragments. It could be of influence that the actors in the theory-condition had the exact same list of adjectives in their instructions as the list that was used in the survey. Some actors were better than others in the sense that they put the stance they intended to show on stage better. Others were better in the sense that the stance they acted was recognized by more spectators. The ratio between validity and inter-annotator agreement differed per actor. It is striking to see that most fragments where the acting was exaggerated were recognized best. For making the virtual suspect this is acceptable as the interrogation game tries to familiarize police trainees with the effects of Leary's theory.

We have seen that the expertise of an actor can influence the perception of his acted behaviour. Fragments of professional actors were rated as more spontaneous than fragments of amateur actors who were rated as more 'acting'. Furthermore, we have seen that the quality of acting (spontaneity) influences the agreement between annotators. Fragments that were rated more spontaneous tended to have lower standard deviations on the ratings of dominance and affiliation. This effect was significant for the deviation of dominance. If one makes an argument for using actors when trying to obtain a ground truth of behaviours, we should be careful to see whether this behaviour is also perceived and interpreted by others to be the intended behaviour. For making a virtual suspect it does not necessarily matter if the actor is experienced or 'good', what matters is that the behaviour is recognized by independent observers. Their judgement should be leading in determining what behaviour to use to model the virtual suspect's behaviour. The perception of the behaviour of this virtual character should then be evaluated. The concern that exists for a human actor also exist for the virtual actor: the intended stance might not be the stance that is perceived by independent observers.

When stills from the best recognized fragments are compared, similarities within stance categories and differences between these categories are apparent, see Figure 4.8. In summary, it can be seen that dominant postures are upright with a gaze straight at the conversational partner while submissive postures are more closed with a gaze away from the conversational partner.

The most valuable lesson learned from these studies is that it is hard to act a stance and -maybe even more valuable- that observers often see diverse aspects in the behaviour of someone. People apparently often show a mixture of stances and it depends on the perspective taken by the judge as to which aspect of the suspect's behaviour determines how his stance is perceived.

5

Social Behaviour in Police Interviews: Relating Data to Theories

The first step towards an artificial agent that can play the role of suspect is knowing what a human suspect does. We analysed a corpus of enacted police interviews to get insight into the social behaviour of interviewees and police officers in this setting[a]*. We collected the terms used to describe the interactions in those interviews. Through factor analysis, we showed that the theories* interpersonal stance*,* face*, and* rapport *and the meta-concepts* information *and* strategy *are necessary to include in a model that captures the social interaction in a police interview. Subsequent validation and relational analysis of the concepts from these theories showed which concepts from these theories are related. This work informed the construction of a virtual agent acting as a suspect in a training game for police officers.*

[a]This chapter is based on: Bruijnes, M., Linssen, J., op den Akker, R., Theune, M., Wapperom, S., Broekema, C., & Heylen, D. (2015). Social behaviour in police interviews: relating data to theories. In *Conflict and Multimodal Communication* (pp. 317-347). Springer International Publishing. [31]

5.1 Introduction

In this chapter, we will describe our investigation of the interaction between police officers and suspects in police interviews by looking closely at what goes on during (practice) police interviews. We have developed a computational model that lets a virtual suspect select the behaviour that is most appropriate for the long-term goal: a virtual suspect that can be used in an application for the training of police students. In this application, the actions of the user are sensed and interpreted to form meaning, for example 'the user is angry'. Our computational model then uses these

interpretations to form a 'mental state' of the virtual suspect, for example 'the police officer is angry and that makes me sad'. The mental state (or mood) of the virtual suspect helps select the most human-like action that the virtual suspect has available, for example 'I am sad so I will make a sad face'. We will not present a completely specified mental model for a virtual suspect in this chapter, but we will provide the groundwork for such a mental model. This chapter will afford two main contributions; first, it will describe how we analysed interpersonal behaviour by validating *ad hoc* interpretations of factors resulting from a factor analysis. Second, we will show which theories from (social) psychology and their underlying concepts were relevant to capture social interactions during police interviews and how these concepts were interrelated.

5.1.1 Data-Driven vs. Theory-Driven

We can follow different paths to build a computational model of virtual suspects and their conversational behaviour in police interviews. One way is to start with a literature study and see what conceptual frameworks in behavioural and cognitive psychology and socio-linguistics—that focus on police interview practice—have to offer. The question is whether these theoretical frames will provide insights to help us in building an operational model for suspect behaviour. Another approach is to start with the application of annotation schemes for specific dimensions of conversational behaviour to perform content analysis of the conversational data. The question then concerns the statistical correlations between aspects of behaviour in this type of data. For example, what is the relation between interpersonal stances that suspect and police take in an interview and the way they manage turn-taking [141] (see also Chapter 3)? The question is whether the concepts (labels) are clear enough and applicable to the data so that the inter-rater agreement is acceptable.

In the current chapter we will take a more holistic approach. The question is what concepts do people use to describe what is going on in a police interview when they experience/observe it. How do they describe the interview and the behaviour of the interlocutors? Does the data, consisting of terms used to describe what is going on in a police interview, reveal interesting patterns? Such patterns might resemble the patterns that theories in behavioural and cognitive psychology and socio-linguistics describe, linking them to the observed practice of police interviewing.

5.1.2 Chapter Outline

We will look at human behaviour in a corpus of police interviews and try to establish which psycho-social theories might explain what happens in these interviews. In Section 5.2, we will give a more detailed overview of our approach. A factor analysis of the occurrence of terms used by observers to describe the interactions in our corpus yielded groupings (factors) of those terms that co-occur. The factors are the basis for a selection of theories and models from (social) psychology, which will be discussed in Section 5.3. In Section 5.4, we will address how the concepts underlying the theories matched the factors used for the factor analysis and how the concepts may be interrelated. To illustrate how these concepts can be used to understand the behaviour

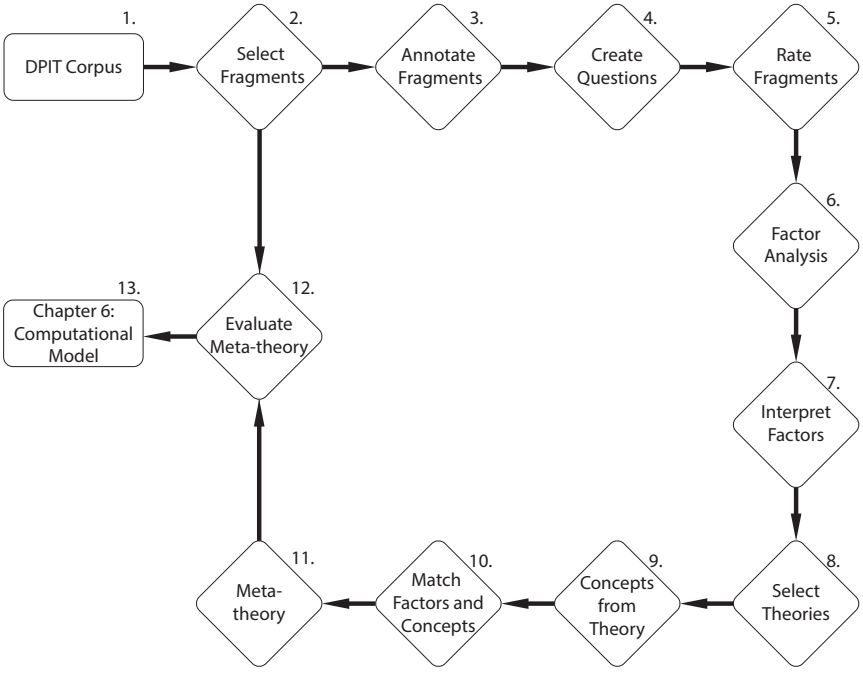

Figure 5.1: Diagram showing the steps taken towards a computational model of police interviews.

of police officers and suspects, we will describe several fragments from our corpus in terms of these concepts (Section 5.5). We will conclude with our thoughts on the creation of a computational model for our virtual agents based on the combination of models and theories (Section 5.6).

5.2 Corpus Analysis

In this section, we will outline how we analysed the behaviours of police officers and interviewees in a corpus of police interviews. We looked at the behaviour of suspects and police officers because we feel modelling the interaction between both parties is necessary to create a believable virtual suspect. In Figure 5.1, we show the steps we took in our analysis.

We started with a corpus of police interviews (step 1), the *Dutch Police Interview Training Corpus* (DPIT Corpus), see Section 5.2.1. From this corpus, six observers independently selected fragments that they thought were 'interesting' in some way (2). For example, these were fragments in which a change in mood or atmosphere took place or fragments in which behaviour could be observed that indicated how the police officers or suspects felt about the interaction. Next, a subgroup of the observers noted as many different *terms* as possible to describe what was going on in the fragments (3)—these could for example be adjectives describing the mood or

nouns indicating behavioural traits. Based on these terms, we created questions with variations of the format *"To what extent is [term] the case?"* (4), see Section 5.2.2. The original six observers then rated a random subset of the same fragments from the corpus on a five-point Likert scale for every question, once for the police interviewer and once for the suspect. For example, a question from the point of view of the suspect was "*To what extent is the suspect dominant?*" (5). We performed a factor analysis to find a clustering of correlated questions (6) which we discuss in Section 5.2.3. Next, a subgroup of the original six observers reached consensus on the interpretation of the factors (7) and we selected well-known psychological and sociological theories that addressed these interpretations (8), see Section 5.3. In Section 5.4, we will discuss how the concepts that these theories employ were matched to the factors by selecting the concepts that fit each factor (9 and 10). This also revealed the relations between the different concepts that were selected (11), as for some factors, concepts from different theories were applicable. This relation between theories allowed us to integrate the different theories into one 'meta-theory' that provides the terms and concepts to describe the interactions in a police interview. We checked whether this 'meta-theory' could describe what is happening in a police interview, see Section 5.5 (12). Our final step (13) was to create a computational model from our 'meta-theory' [29], see Chapter 6.

5.2.1 Corpus Description

The *Dutch Police Interview Training Corpus* (DPIT Corpus) is a corpus that consists of police interviews conducted by trainees of the Dutch Police Academy. For more information about this corpus see chapter 3, section 3.4.1.

5.2.2 Observations of the Corpus

Six observers[1] with previous experience in interaction analysis independently viewed a selection of interviews from the corpus. They selected fragments that they thought were 'interesting in some way' (step 2 in Figure 5.1), for example, fragments in which a change in mood or atmosphere took place or fragments in which specific behaviour could be observed. The observers noted as many different *terms* as possible to describe what was going on in their fragments. The following is an example excerpt from the Remerink scenario (translated from Dutch):

POLICE: My name is Bill [*Surname*][2]. Can we address each other by our first names?

SUSPECT: Well I don't think so.

P: Don't think so? Then I will use Ms Remerink. You can still call me Bill, if you have any questions you can do it like that... eh... eh... Ms Remerink.

One of the observers wrote down:

[1]The first six authors of this Chapter.

[2]Names are fake and/or anonymised for privacy.

> The suspect is *invited to call the police officer Bill*, even though she *insists* that he calls her Ms Remerink. He is trying to *be nice*, but he might give away *power*. Now there is *asymmetry* in the way they *address* each other (officer has to say 'u' while she can say 'je'[3]).

This description provides six descriptive terms for the fragment: *tutoyer*[4], *insist*, *be nice*, *power*, *asymmetry*, and *address*. A subgroup of the observers watched all the interesting fragments and also reported as many terms as possible to describe these fragments using a 'think-aloud' strategy (step 3 in Figure 5.1). In the example above, they added the terms *status* and *cold* to the terms selected by the original observer.

The subgroup of observers added fewer new terms to the entire collection of terms with every successive interview fragment. The first observed fragment yielded over 50 unique terms while only 3 new terms were added to the existing collection after observation of the annotations of the final interview. From this we conclude that we have obtained a sufficiently complete collection of terms necessary to describe the interviews included in our corpus: a *semantic frame* [1]. Eventually, the collection converged on a total of 251 unique terms.

5.2.3 Rating and Factoring Fragments

Based on the semantic frame of 251 terms, we created 227 questions with variations of the format *"To what extent is [term] the case?"* (step 4 from Figure 5.1). We excluded terms that were not suited to create meaningful questions. For example, 'fact' was a term that is too general to yield a sensible question or every question would have to be specific to the scenario. Example questions that were included are *"To what extent is aggressive behaviour the case?"*, *"To what extent is the speaker indifferent?"*, and *"To what extent is there an uncomfortable posture?"*.

The original six observers rated fragments from the corpus on a five-point Likert scale for every question (step 5 from Figure 5.1). The observers scored 14 fragments (with a total running time of 19 minutes) of the corpus on the 227 questions. The rated fragments were randomly selected from the fragments that were selected at step 2 (see Figure 5.1). The fragments were scored by asking the rating questions explained above for both the police officer and the suspect, resulting in every question (corresponding to a term) being scored 28 times.

We performed a factor analysis to find a clustering of correlated questions which indicated which categories of questions—and, by extension, which terms—are related (step 6 from Figure 5.1). Questions that were scored with no variation—that is, they always received the same score—for either the police officer or the suspect were excluded from analysis. This resulted in 9 questions being removed (two were excluded from analysis for the subject, 7 for the police officer). The excluded terms were found to be very role-specific; for example, crying is something the police never does. The factor analyses (extraction method: Principal Component Analysis, rotation method:

[3] In Dutch there is a difference between the second person pronouns 'u' (formal) and 'je' (familiar), both are translated to 'you' in English.

[4] The (French) term for 'to thee and thou', to be familiar, based on the observer's description '[..] invited to call the police officer Bill.'

Table 5.1: The 10 items loading highest on the first suspect factor. This factor was interpreted as *dominant* and *opposed*.

Item	Factor loading
'Building pressure'	.96
'Interruptions'	.95
'Aggressive behaviour'	.94
'Angry behaviour'	.94
'Steering a conversation'	.93
'Accusing the other'	.92
'Attacking behaviour'	.92
'Cutting the other off'	.91
'Worked-up behaviour'	.90
'Raised voice'	.88

Varimax with Kaiser Normalisation) revealed 13 factors for both the suspect and the police, see Tables 5.2 and 5.3.

Based on the related questions, we determined which terms loaded strongly (having a correlation of more than 0.50) per factor. The observers used these terms to interpret the corresponding factors (step 7 from Figure 5.1). For example, the first factor (explaining 19.4% of the variance) for the suspect was interpreted as *dominant* and *opposed*. In Table 5.1, we show only the first 10 (of 54) items with factor loadings for the first factor of the suspect.

A subgroup of four of the original six observers interpreted all factors. The consensus on keywords describing the strongly loading factors of the suspects and the police officers is reported in Tables 5.2 and 5.3 respectively. In general, the observers' interpretations were similar. For example, one of the observers interpreted the first factor of the suspect as 'negative, confrontational and dominant', while another observer interpreted it as describing 'dominant behaviour and frustrations'. Discussing the interpretations among the observers generally resulted in agreement on the meaning of the factors. Some factors (Suspect factors 9, 11, and 13 and Police factor 8) remain unclear as the observers were unable to reach consensus. We attribute this confusion to the few and diverse items that load on these factors.

5.3 Linking Factors to Theories

In this section, we will describe how the interpretation of the factors found in the previous section reflects ideas found in theories from (social) psychology (step 7 from Figure 5.1). Based on the theories discussed in this section, in the following section we will present a meta-theory that describes concepts relevant to the interactions in police interviews. The factors describing interpersonal attitudes will be taken together in Section 5.3.1 on stance; the factors linked to face and politeness will be discussed in Section 5.3.2; the factors linked to rapport will be captured in Section 5.3.3. Additionally, two meta-concepts—*information* and *strategy*—were added to accommodate for the concepts that surfaced in the interpretation of the factor analysis but did not

Table 5.2: Variance explained by each factor for the suspect, with the interpretation of the factors.

Suspect Factor	Variance (%)	Cumulative Variance (%)	Interpretation
1	19.41	19.41	Dominance/Opposed (based on frustration), Strategy/Face (attack)
2	17.40	36.81	Rapport (Building), Together
3	15.59	52.39	Submissive/Opposed, Face
4	6.96	59.35	Together
5	6.65	66.00	Strategy (Annoy)
6	6.19	72.19	Information Exchange (Questions)
7	5.74	77.93	Information Exchange (Lies)
8	5.27	83.20	Strategy (Surround a fact)
9	4.70	87.91	–
10	4.01	91.92	Politeness (Face)
11	3.02	94.93	–
12	2.90	97.83	Rapport (Present)
13	2.17	100.00	– (One item: thank)

Table 5.3: Variance explained by each factor for the police, with the interpretation of the factors.

Police Factor	Variance (%)	Cumulative Variance (%)	Interpretation
1	14.77	14.77	Rapport (missing rapport, negative emotions), Arousal, Opposed
2	13.57	28.34	Rapport (Present), Positivity, Together
3	11.16	39.50	Strategy (Avoid), Information Exchange (Lies)
4	10.46	49.96	Submissive
5	8.83	58.78	Together
6	8.37	67.15	Arousal, Dominance (Competitive), Strategy (Attack)
7	8.01	75.16	Dominance/Opposed (based on strategy)
8	4.61	79.77	–
9	4.56	84.33	Dominance/Together
10	4.52	88.85	Strategy (Confront)
11	4.16	93.01	Strategy (Confront)
12	3.68	96.69	Dominance
13	3.31	100.00	Strategy (Confront)

fit easily in a theory. Factors relating to information exchange will be discussed in Section 5.3.4 and factors linked to strategy will be discussed in Section 5.3.5. In each of the subsections, we will describe the relation between these collections of factors and the theories from (social) psychology, including the concepts underlying those theories (step 8 from Figure 5.1). We will provide examples from the corpus and address work done with virtual agents and the mentioned theories. Also, we will give some examples of systems using the concepts.

5.3.1 Interpersonal Stance

Several interpreted factors for both the suspect (1, 3 and 4 from Table 5.2) and the police (4, 5, 7 and possibly 6 from Table 5.3) are related to the attitude the suspect and the police officer have toward each other. Taken together, these factors sketch the outlines of *Leary's Rose*, a model for interpersonal behaviour [116] (see also Chapter 1, Section 1.3.1.1). Leary's Rose represents such behaviour in categories of interpersonal stance on the dimensions of affect (x-axis) and power (y-axis), see Fig. 5.2A. That is, the underlying concepts of Leary's Rose are part of these axes: the opposing concepts of dominance and submission constitute power, and the opposing concepts of feeling together (positive affect) or feeling opposed (negative affect) constitute affect. The model is often pictured as an ordering of the stances on a circle, situated on the two axes, which is called a circumplex. The circumplex can be divided in eight areas: these are interpersonal stances. The circumplex shows that stances that are close together are more related than those that are further apart on the circle, with opposites being negatively related (Fig. 5.2A). Leary suggests that human stances are affected by the interaction with the conversational partner. This means that two conversational partners influence each other with their stance during a dialogue. Leary calls these interactions 'interpersonal reflexes' and asserts that acts on the dominance dimension are complementary while acts on the affect dimension are symmetric. This means that a dominant act (for example, power display) elicits submissive acts, whereas an act with positive affect (for example, cooperative behaviour) elicits another positive affect act (see Fig. 5.2B). For example, if someone displays dependent behaviour towards another person (submissive and positive), that other person will feel a tendency to adopt a leading stance (dominant and positive) [116].

5.3.1.1 Corpus Examples

In the corpus, we see several examples of different stances. In the Van Bron scenario, the suspect mostly behaves in a detached manner, unwilling to cooperate and expressing this through either competing or defiant behaviour. For example, when Van Bron becomes frustrated about not getting enough time to speak his thoughts and says to one of the officers that they should let him speak, the addressed officer says that he does not need to comply with Van Bron's wishes. As a result, Van Bron becomes somewhat aggressive and acts in a very dominant way, which corresponds to a hostile-dominant stance. On the other hand, the police officer usually displays behaviour with a together stance, for example in the Wassink scenario, in which the police officer does his very best to explain in different words to the interviewee what

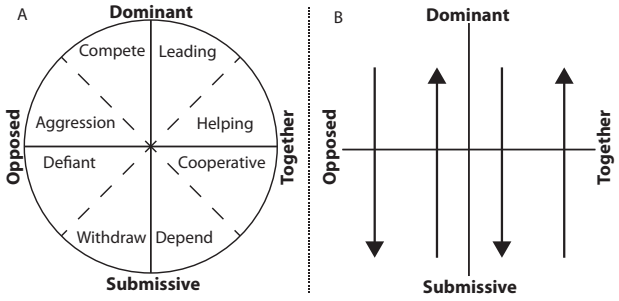

Figure 5.2: A: Leary's Rose is defined by two axes: a dominance axis (vertical), describing the speaker's dominance or submissiveness towards the listener; and an affect axis (horizontal), describing the speaker's willingness to cooperate with the listener. The Rose can be divided into eight areas that each correspond to a stance. B: The solid arrows indicate the behaviour-inviting relation between the quadrants according to Leary's theory [116]. So, dominant-together invites submissive-together behaviour (and vice-versa) and dominant-opposed invites submissive-opposed behaviour (and vice-versa).

he was saying just a moment before. In this attempt to help the interviewee, the officer takes a very positive stance towards her by trying to help and cooperate with her.

5.3.1.2 Systems Using this Concept

There have been a few attempts to create virtual agents that act according to the interpersonal circumplex theory. One of these was the serious game "deLearyous", which focussed on training interpersonal communication skills in a working environment setting, letting users interact with virtual agents through written natural language input [201]. However, one of the findings of this project was that determining the stance of dialogue utterances is a very difficult task, even for human annotators. Other work has focused on finding correspondences of non-verbal behaviour with stances [155]. This approach focusses on the generation of upper body movement and of facial animation on a virtual agent, based on human annotation of behaviours.

5.3.2 Face Threats and Politeness

Informed by Goffman's notion of *face* [70]—a person's public self-image—Brown and Levinson constructed their theory about politeness strategies [28]. Suspect factors 1, 3, and 10 were interpreted as related to face (and politeness), see Table 5.2. The police factors were not interpreted as having a relation with face.[5]

Brown and Levinson distinguish between negative and positive face, which denote a person's need for freedom (*autonomy*) and a person's need to be approved of and approving of others (*approval*), respectively. Their approach to politeness revolves

[5]Police factor 4 was considered by some interpreters to have a relation with face but this was not unanimous.

around the concept of face-threatening acts (FTAs) which are inherent with actions taken by a speaker, as these actions potentially impose on a hearer's face by threatening their needs. Brown and Levinson view politeness strategies as ways to redress these FTAs in order to minimise their imposition. The four main politeness strategy types follow below, ordered from least to most polite.

Bald on-record Being straight to the point, e.g., "*Tell me where you were that night*."

Positive politeness Taking the other's wants into account, e.g., "*Would you like to tell me where you were that night?*"

Negative politeness Not hindering the other's autonomy, e.g., "*If it's not inconvenient to you, could you tell me where you were that night?*"

Off record Being indirect or vague about one's own wants, e.g., "*I don't seem to have written down where you were that night.*"

Conflict situations often arise in the police domain where people may not have the intention to stay polite—on the contrary, they may have the intention to be impolite. Complementary to Brown and Levinson's positive and negative politeness strategies, Culpeper et al. describe impoliteness strategies [51].

Positive impoliteness Damaging the addressee's positive face wants by excluding him or her, being disinterested, disassociating oneself from the addressee or using taboo words. For example, "*Just bloody tell me where you were that night, so I can go home*."

Negative impoliteness Damaging the addressee's negative face wants by being condescending, frightening him or her or invading his or her space. For example, "*Tell me right now where you were that night, or I'll lock you up till Monday*."

5.3.2.1 Corpus Example

On multiple occasions in the van Bron scenario, the suspect demands that the interview takes place according to his wishes. He does this mostly by using sentences that are short and direct, such as "*You have to shut your mouth!*" when he does not receive ample time to speak and expressing his disinterest by replying to the police with short answers ("*It just is.*"). The first utterance is an example of an attack on the police officer's negative face, as the suspect invades his space and claims room for himself in the conversation. The second may not come across as a direct attack on the police officer's face, but it does impose on his positive face, as it indicates that the suspect does not want to cooperate and does not approve of the police officer. Impoliteness is not limited to solely being used by suspects: police officers use impolite utterances as well. This happens frequently when the police confront a suspect with a lie or an incriminating fact. For example, in the Huls scenario the officer is bald on-record and says "*I think you took the money.*"

Even though police interviews can be uncooperative dialogues, politeness is still abundantly used. For example, in the Huls scenario a police officer explicitly expresses

his approval of the suspect's behaviour: "*I think it's decent of you that you try to support your family financially.*" This can be seen as an example of positive politeness, as the police officer takes the suspect's wants (of being approved) into account. In the Motor scenario, an example of negative politeness can be found, as a police officer tries not to impose too much on the suspect's freedom (his autonomy) by saying "*I hope you don't mind too much having this conversation with me.*"

5.3.2.2 Systems Using this Concept

Based on Brown and Levinson's definition of politeness, several systems have incorporated virtual agents that can use utterances that vary in politeness. One of the first of these systems was designed by Walker et al. and involved asking a waiter for drinks with varying degrees of politeness, based on Brown and Levinson's theory [206]. Work by Gupta et al. continued this line of research by creating POLLy, a virtual agent that assisted users in learning English as a second language [80]. This agent took into account how imposing its requests were to the user by redressing these requests according to Brown and Levinson's theory of politeness.

5.3.3 Rapport

The feeling of rapport can be described as being 'in sync' with another person: communication takes place fluently and both interaction partners are roughly on the same level. In our corpus, we see the effects of both the presence and absence of this feeling. Suspect factors 2 and 12 in Table 5.2 and police factors 1 and 2 in Table 5.3 were interpreted as rapport (rapport-like descriptions).

Tickle-Degnen and Rosenthal conceptualised rapport in order to identify nonverbal correlates [193]. Their description of the nature of rapport focuses on the interaction process as a whole and relies on three components of rapport: *mutual attention*, *positivity* and *coordination*. To develop and maintain rapport, interaction partners need to be mutually attentive so that they can achieve a focused and cohesive interaction. Moreover, their interest in the other party should remain at a high level during the course of an interaction. Figure 5.3 shows a schematic view of relative importance of mutual attention and the other two factors of rapport over time. Tickle-Degnen and Rosenthal mention that being positive towards each other is important during the build-up of rapport, yet becomes less important as time passes during interaction. An example of this is language usage among teens, where insults (a sign of low positivity) are the order of the day [207]. Lastly, Tickle-Degnen and Rosenthal describe coordination as having a harmonious relationship between partners—this is the key term related to the feeling of being 'in sync' and is the factor that becomes more important over time.

5.3.3.1 Corpus Examples

In a fragment of the Wassink scenario, the officer and the suspect start to speak more easily and freely to each other after a period of hesitant, slow interaction. The officer starts making gestures and the suspect has her full attention on his comments and responds quickly, without much hesitation. Soon after this, the suspect assumes a

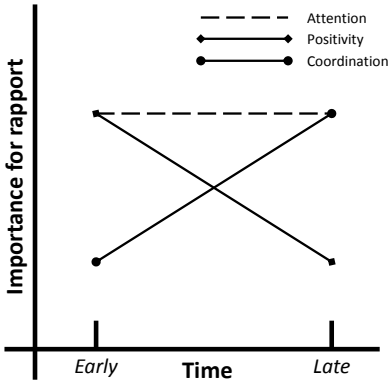

Figure 5.3: Importance of the three components of rapport over time (from [193]).

more interested body posture and finally both parties start laughing together. What is clear here is that the parties have mutual interest in each other and their coordination increases, resulting in them being 'in sync'.

The opposite occurs in a fragment of the Van Bron scenario, in which Van Bron is not listening to what the officers are asking (or does not want to hear what they are saying) and starts making indecent comments about the female officer. In this case, there is little attentiveness of the suspect as well as a lack of intention to be positive towards the officer, resulting in an unpleasant atmosphere in which the police officer is not sure what to say any more.

5.3.3.2 Systems Using this Concept

Huang et al.'s work on the Rapport Agent 2.0—a virtual agent designed to build rapport with users—focuses on backchannelling and turn-taking at the correct moments [90]. Here, backchannelling and turn-taking are used to inform the user of the attention of the virtual agent. Cassell et al. address long-term effects of rapport and how these could be modelled by looking at differences in interactions between friends and strangers [42]. In their research, it became apparent that strangers tend to acknowledge each other more—that is, they make sure that the other party understood that they themselves understood what was being said. Friends are much more direct in their interaction, gazing at each other directly and being less explicit about their understanding of each other, which is explained by them having more rapport.

5.3.4 Information Exchange and Framing

Suspect factors 6 (questions) and 7 (lies) and police factor 3 (lies) were interpreted as having to do with information exchange. The discussion between the interpreters revealed more descriptions of information exchange than just 'lies' and 'questions', but for these other categories no consensus was reached and they are not included in

Tables 5.2 and 5.3. A factor analysis where the questions answered from the points of view of the suspect and the police were taken together revealed more information exchange descriptions during interpretation of these factors, including *give information*, *withhold information*, *lie*, and the notion of *topic or frame*.

Information is exchanged during all conversations between multiple interaction partners. Austin conceptualised an illocutionary act [6] as the intended meaning of an utterance, for example a request for information. Based on this notion, Searle created a classification of five different types of speech acts, namely representatives (informing), directives (requesting), commissives (promising), expressives (expressing a psychological state), and declarations (for official decisions) [173]. The main concepts in this theory are those of requesting information (*questioning*) and *giving* (or not giving) *information*.

A special case of giving or withholding information is *lying*: providing information that one knows or believes to be false. Police officers experienced in interviewing have above average lie detecting skills [125], mainly because they focus on cues that relate to a suspect's story. In other words, inconsistencies in the information exchange are important during police interviews.

The type of information that is exchanged and how it is interpreted is dependent on context—in other words, the interaction's *frame* determines the type of conversation. The notion of frame was first introduced by Bateson in 1955 as he studied the behaviour of monkeys in different situations [15]. Bateson stated that no communication could be interpreted without a meta-message about what was actually going on—that is, what the current frame of the interaction between the monkeys was. During a play frame, all monkeys knew that certain behaviours were accepted (such as biting) which would otherwise be interpreted as a hostile act. Fillmore elaborated on this idea by stating that a frame is "a system of linguistic choices associated with a scene, where a scene is any kind of coherent segment of human actions" [65]. According to Tannen, conversational frames are repositories for social cultural norms of how to conduct different types of conversation, such as storytelling, teasing, and small talk [187]. A frame tells us something about what we can and cannot say in that particular frame. The frame that is currently active allows us to decide which assumptions we can make, customs or 'social scripts' we have (what we can do), and constraints we have (what we should not do).

5.3.4.1 Corpus Examples

Dutch police interviews start with a social frame during which the police officer tries to get to know the suspect and gathers information about the personal life and emotional situation of the suspect. After getting to know each other they continue with a task frame where they discuss the crime that the suspect has been accused of.

Conversational partners do not always agree on the frame that they are using. During the Wassink scenario the suspect does not agree with the social frame the police officer suggests and she asks: "*Why do I have to tell you something about myself?*"

In the Huls scenario the suspect eventually admits to the crime of stealing money from the gas station. During this confession the police officer uses an empathy frame [20] in which he comforts the suspect by telling him that he understands his situation

because he too has children. He agrees with the suspect that it is hard to provide for two children without a stable income.

5.3.4.2 Systems Using this Concept

Multiple virtual agent systems have been created that are at least partly based on speech or dialogue act theories. For example, the Mission Rehearsal Exercise and Stability and Support Operations systems and their derivatives feature agent decision making using speech acts [186]. The same is true of Kopp's virtual museum guide which distinguishes between the performed behaviours and the communicative function of these behaviours [109]. This helps the virtual guide to select responses that vary in their performance, yet have the same communicative function.

Bickmore developed a health counselling agent that bases its reactions on both interpersonal stances and framing [20]. Bickmore uses four different conversational frames to help the agent decide on how to react: the *task* frame, which is used for information exchange; the *social* frame, which is used for social chat and small talk interactions; the *empathy* frame, which is used for comforting interactions; and the *encourage* frame, which is used for coaching, motivating and cheering up interactions. With this information, combined with interpersonal stance, the agent can decide what behaviour to show in different situations.

5.3.5 Strategy Selection

Suspect factors 1, 5, and 8 and police factors 3, 6, 10, 11, and 13 were interpreted as having to do with strategies in interaction, see Tables 5.2 and 5.3. Specifically, the interpreters used the concepts *confront, surround, evade*, and *annoy*.

During communication, individuals make use of strategies to achieve their desired goals. These strategies play an important role, especially during non-cooperative communication such as in the police domain, as described in Chapter 1. Traum et al. [195] describe a set of negotiation strategies—including finding the issue, attacking to aggressively attain a goal, and advocating or proposing solutions—and assert that the negotiating party must balance three goals to be successful in a difficult negotiation. The negotiator has to find an acceptable solution for the problem, gain and maintain the trust of the other participant(s) and manage the interaction by setting the agenda and controlling the topic. According to Campos et al. [40] conflicts are akin to story-plots. They identified components of conflicts that are similar to the components of a plot in a story: conflicts have participants, a cause, and an initiating action, to which responses are made that escalate the conflict, until the conflict reaches a climax after which the conflict de-escalates until there is an outcome that resolves the conflict. Thomas [190] argues that people can take five 'conflict-handling modes' to resolve conflicts: *accommodation*, *avoidance*, *competition*, *collaboration* and *compromise*. These are classified on two underlying dimensions: assertiveness and cooperativeness. The *Table of 10* by Giebels [66] describes the strategies a police officer can use when, for example, they want to convince the suspect that cooperation will be of mutual benefit (see Table 5.4).

Table 5.4: Table of 10 by Giebels (translated from Dutch from [202]).

#	*Strategy*	**Principle**	**Description**
1	*Be Nice*	Sympathy	Show willingness to talk, react empathic.
2	*Be Equal*	Equality	Emphasize commonalities, name external foes.
3	*Be Credible*	Authority	Show trustworthiness, show expertise.
4	*Emotional Appeal*	Self-Perception	Play on feelings (consider victims), offer to earn respect.
5	*Intimidation*	Insecurity	Warn of consequences, personal attack.
6	*Impose Boundaries*	Scarcity	Deny concessions, ignore opponent.
7	*Direct Pressure*	Repetition	Repeat appeal (plant seed), accomplished fact.
8	*Legitimate*	Legitimacy	Refer to rules and laws, refer to other opinions.
9	*Trade*	Mutuality	Ask for something in return, concession after high commitment.
10	*Convince Rationally*	Consistency	Bring forward arguments, confront with contradictions.

5.3.5.1 Corpus Examples

The Remerink scenario in our corpus starts with a frustrated suspect who is apparently angry about something. The police officer uses a negotiation strategy to find out what is bothering her in order to resolve this issue. He asks the woman what is bothering her and eventually she says she is angry about the method with which she was picked up from her house. She is ashamed and angry about the way they came to her house and brought her in with all the neighbours watching.

Later in the Remerink interview, the suspect is accused of taking money from her ex-husband and she becomes emotional and silent every time when the topic of her husband comes around. Due to the fact that the topic is undesirable for her to talk about she tries to avoid going into it any further.

In the Huls scenario, the police officer is surrounding a specific fact during the conversation, so the topic cannot be avoided. He continues to aggressively put similar questions to the subject to put pressure on him to tell the truth.

5.3.5.2 Systems Using this Concept

The Mission Rehearsal Exercise and Stability and Support Operations systems and their derivatives feature virtual agents in war scenarios with which users have to negotiate [186]. These scenarios deal with dilemmas the user has to solve. For example, a user has to convince a local Afghan doctor to move his clinic to another location as the user has to conduct military operations in the area of the clinic. One of the ways to convince the doctor is using rational arguments such as offering incentives. Fur-

Table 5.5: Concepts within the theories *stance*, *face*, and *rapport* and the meta-concepts *information* and *strategy*.

Stance	Face	Rapport	Information	Strategy
Friendly (Dominant-Together)	Autonomy+	Coordination	Questioning	Confront
Aggressive (Dominant-Opposed)	Approval+	Attention	Give info	Surround
Withdrawn (Submissive-Opposed)	Autonomy–	Positivity	Lie	Evade
Dependent (Submissive-Together)	Approval–		Withhold info	Annoy
			Frame/topic	

thermore, this system also takes emotional consequences into account when deciding whether to cooperate with or oppose the user.

5.4 Relations between Factors, Theories and Concepts

In the previous section, we discussed what theories from (social) psychology match the interpretations of the factors found in the factor analyses (see Section 5.2.3) and we explained the concepts from these theories (step 9 from Figure 5.1). In this section, we will discuss how these concepts are related to the factors (step 10 from Figure 5.1). This will give insight into both the relations between the factors and the concepts, and the relations between the concepts themselves (step 11 from Figure 5.1). Based on our findings on these relations, we will describe how the theories from which these concepts originate are connected.

5.4.1 Concepts in Theories

Psychological and sociological theories use concepts and describe the relations between these concepts. Theories provide us with a way to describe an interaction (in our case, a police interview) and they can be used by a virtual tutoring agent (in our case, a virtual suspect) to predict the effects of its behaviour in an interaction with a human user. For example, the central concepts in the interpersonal stance theory are *dominant*, *submissive*, *together*, and *opposed* and the theory describes how the combination of these concepts creates the notion of 'stance' and predicts how people are influenced by the stance of others. A virtual tutoring agent can use this knowledge to create an interesting and useful learning experience. For example, a user might learn by experiencing that if he displays opposed behaviour in an interrogation the conflict escalates. The virtual suspect can display opposed behaviour in an attempt to get the user to also display opposed behaviour and then let the conversation escalate [30]. Each of the theories we selected in the previous section has such concepts, see Table 5.5. The concepts from face are positive and negative *autonomy*, and positive and negative *approval*. The concepts for rapport are *coordination*, *attention*, and *positivity*. We added two meta-concepts—*information* and *strategy* to accommodate for the concepts that surfaced in the factor analysis interpretation but did not fit easily in a theory. The information concepts we found are *questioning*, *give information*, *withhold information*, *lie*, and *frame* or *topic*. The strategy concepts are *confront*, *surround*, *evade*, and *annoy*.

5.4.2 Factors: Theories and Concepts

The interpretation of the factors, see Section 5.2.3 and Tables 5.2 and 5.3, and the matching of these factors to theories leads to links between theories and factors. To validate these links, four observers indicated with which concept(s) from the theories (Table 5.5) a factor could be explained. This method provided us with a possibility to validate the intuitive (subjective) interpretation of the factors that is common practice in the field of social science. In other words, we used the initial interpretations to select theories that 'cover' the factors and we used the concepts from these theories to validate the labels for the factors. This matching of factors to concepts is a data-driven interpretation of the factors (see Section 5.1.1) and might bring us closer to a 'correct' interpretation of a factor.

In Table 5.6, we show the cumulative score the observers gave the concepts for each factor. The colour coding in this table indicates how much the observers agreed that the concept could explain the factor: the dark-coloured cells indicate unanimous agreement, the light-coloured cells indicate that three out of the four observers agreed. The initial factor analysis interpretation of the factors is indicated with an asterisk.

The fit of the factors and the concepts determines the validity of the interpreted theory for this factor. The observers unanimously matched most factors to the concepts corresponding with the initial factor analysis interpretation of the factor; see Table 5.6 in which the asterisks indicate the initial interpretation and the dark cells indicate unanimous matching. The factors where the observers disagreed (not unanimous) with the initial factor interpretation are police factors 1, 9, 11, and 13 and suspect factors 1, 3, and 10 (see Table 5.6).

We see several explanations for this disagreement; first, the initial subjective interpretation of the factors might have been wrong. The factors with a higher number had fewer items loading on them (and less explanatory power), which might have made it more difficult to interpret them. Four of these higher factors (suspect factors 9, 11, and 13 and police factor 8) had few and diverse items loading on them, which resulted in disagreements during the initial factor interpretation. It is likely that the disagreement persisted in the current analysis for suspect factors 9, 11, 13 and police factor 10. Second, factors could initially have been interpreted as having an 'absence of something'. This was the case for police factor 1 which was initially interpreted as 'missing rapport'. In the subsequent 'mapping concepts to factors' task, the observers did not unanimously match the concepts of rapport to police factor 1. This might be because the instructions were unclear what to do when a concept was explicitly absent: some observers said that this factor contains information about the concept rapport (i.e. rapport is missing) while others said there is no rapport so the concept of rapport is not present. For suspect factor 3 and 10 we can give no alternative explanation and conclude that our initial factor interpretation was incorrect (see Table 5.7).

5.4.3 Relations between Theories

The observers unanimously matched factors to concepts from theories that were not initially included in the factor analysis. In other words, more concepts than initially came to mind might play a role in explaining a factor. For example, suspect factor 4 was interpreted during factor analysis as a *together* stance, but the factor was not only matched to the concept *together* (from theory of stance), but also to *positive approval* (from the theory of face), and *attention* and *positivity* (from rapport) (see Table 5.6).

Our methodology makes clear how the theories are related to each other. For each factor, concepts from different theories can be applicable. This *co-occurrence of concepts* suggests that the corresponding theories are related. In Table 5.8 we show in how many factors concepts co-occur. For example, the interpersonal stance *together* co-occurs in more than one factor to: *positive approval*, a concept underlying face (4 co-occurrences), and *coordination* (2), *attention* (4), and *positivity* (5), which underlie rapport (see Table 5.8). This is indicative of a strong link between these concepts. The concept of *dominance* co-occurs with the concepts: *opposed stance*, *negative autonomy*, *negative approval*, *confrontation*, and *annoy* (see Table 5.8). To investigate this relation further, we look at Table 5.6, which shows that suspect factor 1 was matched with the concepts dominance, opposed, negative approval, and annoy. For the police factors dominance co-occurs with negative autonomy and confront. This is likely due to the different roles the interactants have. The police officer assumes a dominant stance when he confronts the suspect with an incriminating fact: the act of confronting is dominant. This might be strengthened by the power the police officer has, as he dictates the course of the interview: a concern for the autonomy of the suspect. A dominant suspect might use the strategy annoy to intentionally thwart the progress of an interview and this could negatively impact the approval of the officer. Linssen et al. [122] proposed that interpersonal stance and politeness (face) are related. They suggest that the dimensions of power and affect used in the model of Leary's Rose can be mapped to the dimensions of face: autonomy and approval. For example, when a person is very dominant, she does not take the other's autonomy into account. A similar relation holds for the dimension of affect, as a person who is opposed to someone else expresses disapproval of that person.

We will further investigate the relations between the different concepts in the next section and illustrate them using examples from the corpus. The related theories were integrated to form the basis for a computational model. As each theory describes relations between the cause and effect of behaviour in an interaction, a virtual tutoring agent (virtual suspect) can use a computational model of these theories to predict the effects of its behaviour in an interaction with a human user (see Section 5.6 on future work).

5.5 Illustration of Relations

In the previous section, we showed that certain concepts underlying the theories appear to be related based on the data from our corpus (see Tables 5.6 and 5.8). Here, we will illustrate several of these links with example fragments from the corpus that were not used for annotation and the subsequent factor analysis. We will illustrate

Table 5.6: The matching of concepts from the models with the factors derived from the clustering of terms. The numbers indicate how many (out of 4) observers thought the concept from the theory (column) fit the factor (row). Asterisks denote the initial interpretation of the corresponding factor.

Factor		Stance				Face				Rapport			Information					Strategy			
		Dom	Sub	Tog	Opp	Aut+	App+	Aut–	App–	Coor	Att	Pos	Quest	G. Inf	Lie	W. Inf	Frame	Confr	Sur	Evade	Annoy
P	1	2			4*			2	4	2*	1*	2*	2			1	1	1		2	4
	2			4*			4			2*	4*	4*		1			2				
	3	1		1	3			1	3			1	1		4*	4	3		1	3*	
	4		4*		1	1			2	1	2	2		1		2				3	
	5	2	1	4*		2	4	1		3	3	4	1				1		1		
	6	4*		1	3*			4	2	1		1	1	2			1	4*			
	7	3*	1		4*	1		2	4			1	1			1	1	2		1	1
	8							3		1	1		1						1	1	
	9	4*		1*			2	2			1	2							1		
	10							2	1				1	1			1	3*	3		
	11	1	1		2			2	1									2*		1	1
	12	4*		1	2		1	3	3	2	1	1	3	1				2			
	13	1	1	3		3	2			1		3		2	1			1*			
S	1	4*			4*	*	*	3*	4*	1	2	1			1	1		3		2	4
	2			4*		3	4			4*	4*	4*	1	1			1				
	3		4*		1*	1*	*	2*	3*			1	1	1		2	2	2		3	
	4	1		4*		2	4			3	4	4		2							
	5	2	1		4	1		2	4	1	1	1	1			2	3	1		1	4*
	6	2		1				1		2	3		4*	3				1	2		
	7		1		2	3	1	1	2	1	1	1	1		3*	3	1			4	
	8	1	1	1		1	3	2	1	2	1		2	1		2		1	4*	2	
	9	1	1	2		3	1	2	1		3	1		2			1	1			
	10			2		3*	3*	1*	1*	2	2	2		1			1				
	11	2	1					3	1	1	2	1		1				1	1		
	12		1	4		2	3			4*	4*	4*									
	13			1								1									

Table 5.7: The factors from the factor analysis with the concepts that were unanimously matched to the factors. A dash means no concepts were unanimously matched to the factor.

	Factor	Interpretation based on concepts
Police	1	Opposed, Negative Approval, Annoy
	2	Together, Positive Approval, Attention, Positivity
	3	Lie, Withhold Info
	4	Submissive
	5	Together, Positive Approval, Positivity
	6	Dominant, Negative Autonomy, Confront
	7	Opposed, Negative Approval
	8	-
	9	Dominant
	10	-
	11	-
	12	Dominant
	13	-
Suspect	1	Dominant, Opposed, Negative Approval, Annoy
	2	Together, Positive Approval, Coordination, Attention, Positivity
	3	Submissive
	4	Together, Positive Approval, Attention, Positivity
	5	Opposed, Negative Approval, Annoy
	6	Questioning
	7	Evade
	8	Surround
	9	-
	10	-
	11	-
	12	Together, Coordination, Attention, Positivity
	13	-

Table 5.8: The observed relations between the theories based on the co-occurrence of their concepts in the factors. The numbers indicate how often these concepts co-occurred. Only the concepts rated unanimously present in the factors are shown.

		Stance			Face		Rapport			Info		Strategy	
		Dom.	Tog.	Opp.	App+	App–	Coor.	Att.	Pos.	Lie	W. info	Confront	Annoy
Stance	**Dominant**	x		1		1						1	1
	Together		x		4		2	4	5				
	Opposed	1		x		4							3
Face	**App+**		4		x		1	3	4				
	App–	1		4		x							3
Rapport	**Coordination**		2		1		x	2	2				
	Attention		4		3		2	x	4				
	Positivity		5		4		2	4	x				
Info	**Lie**									x	1		
	Withhold info									1	x		
Strategy	**Confront**	1										x	
	Annoy	1		3		3							x

the *co-occurrence of concepts* (see previous Section) that shows the relation between the concepts of different theories. Also, we will illustrate how our findings might be extended to explain the dynamic aspects in a police interview.

5.5.1 Co-occurrence of Concepts

We found the strongest links between the *together* stance and *positive approval* concepts and between the *opposed* stance and *negative approval* concepts (see Table 5.8). An example from the Bruintjes scenario (see transcript below) shows a together stance occurring together with positive approval. In this fragment, the police officers are asking questions about the suspect's leisure time, to which the suspect responds that she spends most of her time at the mall with her girlfriends. The police officers respond to this by indicating that they understand what she means ("*Just chilling.*") and they all start laughing about this. In this moment, the police officers are very much trying to sympathise with the suspect, thereby adopting a together stance. They are also expressing approval by saying that they understand the suspect's wish to stay at the mall. In the preceding section, we also showed that there is a strong link between the together stance, concepts underlying rapport (particularly attention and positivity) and positive approval. In the corpus fragment about the suspect staying at the mall, it is clear that both interaction parties are paying a high degree of attention to each other. One of the police officers is asking questions about the suspect's activities which yield immediate responses from the suspect. There is, however, no uncomfortable atmosphere during this part of the conversation, as both the officers and the suspect start laughing about this topic. Thus, the concepts of a together stance, positive approval, and both attention and positivity are displayed in this part of the conversation.

POLICE OFFICER: Those girlfriends, eh, 'cause you said you go shopping with your girlfriends...
SUSPECT: Mm mm. [*Confirmatory.*]
P: Do you have good friends? Close friends?
S: [*Nods enthusiastically.*] Yes.
P: Yes?
S: Yes.
P: So what do you go and do with your friends?
S: Yeah, well, basically, we are often at the mall.
P: At the mall?
S: Yeah, one of those indoor malls.
P: And what do you do there?
S: [*Softly:*] Kind of hanging around. [*Laughs.*]
P: [*Laughs.*] Just chilling.
S: [*Laughs.*] Yeah!

An example of the strong link between an opposed stance and negative approval can be found in the Wassink scenario (see transcript below). In this excerpt, Mrs Wassink, the suspect, is asked whether she wants to cooperate with the police officer by

answering some of his questions, because he wants to form a picture of her situation. Mrs Wassink does not comply and indicates that see does not see the point of doing so.

POLICE OFFICER: I don't know you and you don't know me either.

SUSPECT: No.

P: But maybe it would be convenient if we would first discuss some things about you—about who you actually are. Do you think that's OK?

S: Well... Why?

P: You don't think that's useful?

S: [*Shrugs, shakes her head.*] No, I don't know why I should tell you who I am.

P: Yeah. [*Short pause.*] Well, I would like to know.

S: But what for?

P: Because I would like to know who I'm talking to before I can talk with you—what you just indicated, that you might have physically abused someone.

S: [*Stares blankly.*]

P: Do you understand what I'm saying?

S: Yes, I do, but that you would like it, well... [*Shrugs.*] Are you going to tell me something about yourself as well or...? [*Shakes head.*]

P: Well, I don't know, would you be interested in what I would have to say?

S: No, but I also don't understand why you would be interested in me.

In the above fragment, the suspect gives snappy replies to the police officer's questions. This is a first indication that she is not trying to cooperate: she seems to have a very opposed stance to the officer, his approach and his proposals. With her behaviour, the suspect also expresses disapproval (negative approval), repeatedly shaking her head and shrugging, indicating that she does not agree with the police officer or just does not care. Moreover, Mrs Wassink goes a bit beyond simply disagreeing, as she seems to intentionally annoy the police officer. She does this by questioning the police officer's approach (repeatedly asking of what use it is), by expressing that she does not understand what is going on, and by asking a counter-question (whether the police officer will say something about himself as well). Here we see that an opposed stance, negative approval, and an annoy strategy occur together.

We found several other relations between concepts from the theories we used that occurred less frequently in our collection of factors. On some occasions, the police officer, but mostly the suspect, used a confront strategy which was accompanied by a dominant stance and negative approval. In these cases, the suspect was trying to lead the conversation by confronting the police officer(s) with his or her own opinions (which were negative in nature most of the time). Another striking co-occurrence of concepts is that of the concepts underlying rapport. As Tickle-Degnen and Rosenthal assert [193], coordination, attention and positivity generally occur together to form the feeling of rapport and this is confirmed by our observations.

5.5.2 Concept Dynamics

Our approach to analysing the corpus of police interviews hinges on the annotation of short fragments. However, our annotation did not capture the dynamic aspects of the interviews, for example how and why people change stances or how their feelings of rapport increase or decrease. Here, we illustrate how this may work by describing a change in a situation in the Van Bron scenario (see transcript below) in terms of the concepts from Table 5.8. In this case, the suspect is asking the police officers their name and surname in a dominant way. One of the police officers immediately agrees to give his surname, but the other only gives his first name. To this the suspect replies by making a small gesture with his hand, implying that he also wants to know the officer's surname. The officer then responds in a laconic way by saying "*Oh, you want my **surname**?*" which leads to the suspect imitating the officer's response and adding "*Wise guy.*" What we witness here is an exchange in which the police officer has an opposed stance towards the (dominant) suspect and acts disapprovingly of the suspect's behaviour by not granting a full answer. This, in turn, elicits a similarly disapproving response by the suspect. Thus, over time, the suspect changes from having a dominant stance towards one that is more opposed to the officer because of the latter's behaviour.

SUSPECT: [*Points at police officer 1.*] What was your name again?

POLICE OFFICER 1: **** [*Gives his first name.*]

S: And...? [*Makes a gesture with his hand for the police officer to complete his name.*]

P_1: **** [*Gives his surname.*]

S: ****? [*Repeats the name.*]

P_1: Yeah.

S: [*Chuckles.*] And you? [*Points at police officer 2.*]

POLICE OFFICER 2: **** [*Gives his first name.*]

S: [*Makes the 'completion' gesture again.*]

P_2: What is that? [*Mimics the gesture.*]

S: [*To P_1:*] He doesn't understand? **** [*P_1's surname.*] And you? [*Points at P_2.*]

P_2: Ah, you want to know my last name?

S: [*Stares at P_2.*]

P_2: Yeah, if you could just be clear in your questions...

S: He understands [*Points at P_1.*] are you a bit stupid or something like that?

P_2: Yeah, I'm a bit more stupid than him, OK...

S: That's clear.

P_2: [*Softly:*] All right.

S: No, it's not all right. What is your surname?

P_2: **** [*Gives his surname.*]

S: Ah.... **** [*Repeats P_2's surname.*] [*Softly:*] Wise guy.

5.6 Conclusions

In this chapter, we presented our methodology for analysing the behaviour of police officers and suspects in a corpus of enacted police interviews. Taking a holistic approach, we described fragments of this corpus in short terms that captured the behaviour of the participants in, and the atmosphere of, the interviews. We used a factor analysis to cluster related terms based on ratings of observers who annotated to what extent these terms were applicable to each fragment. Based on the factors we found, we selected theories from (social) psychology that we intuitively thought could explain these factors. We included theories about interpersonal stance, face and politeness, and rapport and defined two meta-concepts, namely 'information' and 'strategy', to account for the interpretations that remained. To determine whether these theories matched the factors, we investigated whether these factors could be explained by the concepts underlying the theories. We found that our initial factor interpretation and the match between factors and concepts overlapped broadly. We also found that many factors were matched to more concepts than initially were associated with the factors. We used this finding to create a collection of interrelated concepts that gives insight into how the different theories relate to each other. With this collection, we are able to (at least partly) describe the behaviour of both police officers and suspects in an interview setting.

Our combination of holistic and theory-driven methodology does, however, have its limitations. As is the case in most observational studies, our annotations of the police interview corpus were based on our interpretations of the behaviour of the interacting parties and thus subjective. For future work, our methodology may be repeated to include more observers (and more independent observers) which may lead to a broader semantic frame, possibly alleviating problems inherent with interpretation of behaviour.

A second limitation of our approach is that it currently focuses on describing *short* fragments from the corpus. In the previous section however, we illustrated how our findings may be extended to explain changes in the behaviour of interacting parties over longer periods of time. We based these examples on how temporal aspects are explained by the theories from which we drew our concepts. We wish to continue this line of research by investigating how the interplay of these concepts influences the dynamics in police interviews. This may for instance be done by locating moments in our corpus in which a person's behaviour changes. For example, there may be moments when a person changes his or her stance or becomes less polite. Someone may also consciously change his behaviour to evoke desired behaviour of the other party. This may for example be the case when a police officer adopts a 'together' stance to build rapport with a suspect. Thus, the communicative contexts before and after this type of changes in behaviour should be compared, to discover what may have caused the change in behaviour. This causality is of vital importance for the creation of a virtual suspect agent, as such an agent needs to be capable of taking logical (and explainable) actions.

This calls for an extended empirical study of the corpus, which remains future work. Such a study may also validate the links we found between concepts, as our

current work only investigated a number of fragments from our corpus. Lastly, we also wish to investigate how the methodology we present in this chapter translates to other domains. Whether our approach can be used to analyse communicative behaviour in other domains depends on the availability of a corpus and theories on interaction that explain (parts of) the behaviour. A related domain in which we are also involved is that of street interventions by police officers with loitering juveniles (e.g. [120]). This domain features a different setting and the environment imposes other restrictions on the interaction, such as an easier 'way out' for the juveniles because they are not kept in a room like the interviewees. Still, this domain does not differ strongly from the domain discussed in this chapter, as they are both related to police work and display the unique features this work has such as the status of the police officer. It remains to be investigated whether our approach would allow for feasible analysis of behaviour in a completely different domain. However, given a sufficiently rich corpus of such interactions, we expect that our methodology can be used to analyse corpora from other domains as well.

The final consideration on the work in this chapter is that we had access only to actors who played the role of a suspect and not to 'real' suspects. These actors are professional training actors and they are well versed in training the students of the police academy through role-play, yet they remain actors. Not all actors are created equal and there might be actors or 'role-plays' that do not fully resemble a 'real interrogation', see also the previous chapter. Our quest is to create virtual characters that show 'real' and 'human-like' behaviour, so basing the behaviour of such virtual characters on behavioural data of actors seems somewhat counter productive. However, our virtual training actor will have the same role as the human actors their behaviour is based on and the behaviour of these human actors is sufficiently 'realistic' for training.

In chapter 6, we will construct a mental model for virtual agents that play the role of a suspect in a police interview setting. As indicated above, we will focus on the dynamics of such interviews, establishing a computational model that enables a virtual agent to perform causal reasoning. This system will go beyond being an 'autonomous sensitive artificial listener' as in [171]. The system will be able to use the 'mood' of its mental model to select the most appropriate action it has available. The work in this chapter informed the creation of this model, which in turn will be used for virtual agents in a tutoring application.

Part III

Suspect Response Selection Model

Part III: Suspect Response Selection Model

In this part of the thesis, I will present our work on creating and evaluating a response model for a virtual character that will play the role of suspect in a practise police interrogation. The response model will determine the types of behaviour that the virtual character will give in response to the questions posed by the police student. In chapter 6 I will discuss the creation of the virtual suspect's response model. In chapter 7 I will discuss the evaluation of this response model. We considered how to evaluate something as abstract as a response model. We evaluated how well the response model is able to show behaviour that interactants can identify as belonging to a persona. A virtual suspect-system will need to respond as 'human-like' as possible. Such a system is a combination of components that each have their own level of 'human-likeness'. We evaluated the response model in two settings: separate from other components and in a complete virtual suspect-system.

6

A Virtual Suspect Agent's Response Model

We developed a computational interpersonal affective response model for virtual characters that act as suspect in a serious game for training interviewing (interrogation) skills to police officers. We implemented a model that calculates the responses of the virtual suspect based on psychological and sociological theories and observation of (practise) police interrogations. We will describe the aspects of the move (question asked) by the police interviewer that we will distinguish and how the suspect responds to the move. This response is dependent on personality characteristics of the suspect character (persona) and on the dynamic state of the interaction. In this chapter we will present our response model for such a virtual suspect agent.[a]

[a]This chapter is based on several publications [31, 34, 35, 208]. The main publication is: M. Bruijnes, S. Wapperom, R. op den Akker and D.K.J. Heylen. *A Virtual Suspect Agent's Response Model.* in Fourteenth International Conference on Intelligent Virtual Agents (IVA 2014); Proceedings of the Workshop on Affective Agents, L. Ring, Y. Leite and J. Dias (eds), Gaips Intelligent Agents and Synthetic Characters Group, Porto Salvo, Portugal, pp. 17-24, 2014 .

6.1 Introduction

Understanding the dynamics of social relationships is in many professions an important skill and this skill is often trained using role-play. Role-play with an actor can illustrate the effects that the social behaviour of the student has on the social behaviour of the actor and vice versa. A serious game with a virtual character can also successfully train social skills to individuals. Users in a virtual environment tend to perceive virtual characters as actual humans, and respond to them in a naturalistic way regarding affect, personal space, and social presence [24]. Thus, a virtual agent can play the antagonist in human-agent role-play, which in our case means that the

virtual human plays the role of suspect in a training for police officers to hone their interrogation skills.

A virtual suspect needs to respond to the police officer as a human suspect would, for example if the suspect is being treated in an unfriendly manner, the suspect will respond angrily. A way to look at this is that a virtual agent needs three main capabilities to be able to have a meaningful *social* interaction with a user. These three capabilities are often characterized as: *sense, think, act* (e.g. [174]). The actions of the user have to be *sensed* and interpreted (e.g. the user says "*Tell me where you were last night, scum!*" which is interpreted as a dominant and aggressive request for information). The virtual human system has to *reason* (or 'think') about the interpretation of the user's utterance (e.g. the user is dominant and aggressive which makes me angry and non-compliant so I will not give any useful information). The response should take into account the specific role that the agent plays: in this case a suspect. The role determines what tactics and psychological manoeuvring is involved and this should be reflected in the response [31, 138]. A response based on human behaviour can be used to make the behaviour of the virtual human more believable to humans [179]. Based on such reasoning the virtual agent system can select the most appropriate behaviour in its repertoire and display it as a response (e.g. according to my reasoning I am now sad and angry, so I will make a sad face and say "*You're not nice!*"). The human responds to this *act* of the virtual human and the cycle continues.

We propose a model for the reasoning about the dynamics of social relationships in a police interrogation from the perspective of a suspect. Our model can select an appropriate response for a virtual suspect based on the interpersonal stance of the interrogator and the personality of the suspect: we call this a *Response Model* (RM).

6.1.1 A Serious Game

In the serious game for social skills training that we envisioned, a virtual human playing the role of suspect would respond to the verbal statements or questions of a user that plays the role of a police officer. Before the role-played interrogation, the user is presented with a scenario akin to a police report which includes information about the suspect and what questions to ask. This scenario also contains the goal of the role-play from the perspective of a game, for example the game objective could be to get the suspect to give information about a certain topic. Additionally, the serious game has learning objectives that are tailored to the user. For example, the user has to learn how to maintain a friendly stance with an obnoxious suspect. Here the game could set the learning-goal 'to remain friendly' and provide a challenge by creating a virtual suspect that has an obnoxious personality (see Linssen et al. [120]). The interview starts with the user asking a question or giving a statement for which the user can use his or her own words. The system can gather this free text from a speech signal by automatic speech recognition or by having the user type in a chat interface. The virtual suspect can give responses that include or exclude the information that is sought by the user, for instance by telling the truth or lying[1]. All the responses available to the virtual suspect are held in an information state.

[1]This does not mean that the suspect is always guilty. An innocent suspect can also lie or omit to tell the truth, for instance because he or she does not trust the police.

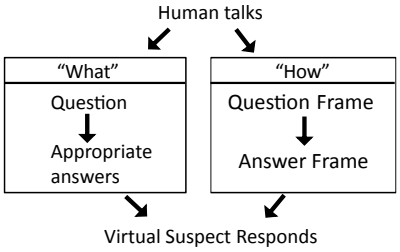

Figure 6.1: A virtual suspect agent needs two things in order to respond appropriately: understand *what* is said and *how* it is said. The *what* deals with the content of the question a user asks and comes up with appropriate answers, for instance through Q&A matching. The *how* deals with how the user asks questions and *how* the virtual human should respond. Our response model deals with the 'how'.

The RM we propose in this chapter determines the type of response that would be appropriate from the perspective of the social relation the user has with the virtual suspect. This includes, on an abstract level, what information the suspect provides. The interrogation is completed when the user meets his or her objectives or is no longer able to attain a goal. Feedback about the user's performance can be presented during the interrogation, for example in thought bubbles that disclose the virtual suspect's thoughts and feelings [121]. Additional feedback can be presented after the interrogation is over to give the user the opportunity to learn by reflecting on his or her interaction [27].

Sacks, Schegloff, and Jefferson [165] suggested turn-taking is systematic and neatly organised 'turn-by-turn', yet in chapter 3 we saw that in a police interview this is not always the case. We found that occurrences of overlap and silences are common and that turn-taking can carry meaning. Ideally, for a serious game the turn-taking between the user and the virtual suspect will be human-like, meaning that the interlocutors can interrupt each other and talk at the same time. The focus of our work was on modelling realistic and recognizable interpersonal relations for a virtual suspect. We did not tackle the (technical) challenges of building a virtual human that can engage in real-time turn-taking, see for examples of real-time fluid turn-taking systems [55, 137, 191]. Our prototype, that will be discussed in chapter 7, offers only turn by turn interactions.

In order to respond appropriately and remain credible, the virtual suspect agent needs capabilities that can be split in two ways: understand 'what' is said and 'how' it is said. We split these two information flows, see Figure 6.1. The 'what' deals with the content. The system can propose a set of appropriate responses based on what the user said through for example Q&A matching (e.g. [75, 118]). To the question *"What time is it please?"* the system could have several answers, for example a friendly *"It is 12 o'clock"* or an unfriendly *"Buy your own watch"*. 'How' the user asked the question needs to influence the answer of the system to remain emotionally credible. Based on how the user asked the question the most appropriate answer can be selected, in this example the friendly answer seems likely. This interpersonal 'colouring' of the interaction is what Walker et al. [206] called Linguistic Style Improvisation. We focus

on the 'how'-part of the interaction. Our RM works independently from components in a virtual human dialogue system that, for instance, automatically classify a user's utterance, compute the content of the message, or manage the turn-taking of a virtual agent.

Our analysis of the DPIT-corpus (see chapter 5) showed us *how* suspects and police officers interact. In particular it gave us insight into the social behaviour of police officers and suspects in the police interview setting. We collected many terms that people use to describe the interactions and relations in the corpus. A factor analysis revealed factors that could be interpreted in relation to the theories of *interpersonal stance* [116], *face* [28], and *rapport* [193] and the meta-concepts *information* and *strategy*. These theories provide a way to describe the interaction in a police interview. Each of these theories and meta-concepts is a collection of concepts (see Table 5.5) and all these concepts are used to describe the social relations within police interviews. Therefore, we argue that these concepts must be included in a response model for a virtual suspect if the model is to capture the dynamics of the social relations and interaction of a suspect and a police officer in a police interview. The rest of the chapter will deal with the motivations for design decisions (section 6.2) and the implementation of our RM (section 6.3). In chapter 7 we will describe a prototype where the RM is integrated in an embodied virtual human dialogue system.

6.2 An Overview of the Response Model

The RM consists of four components: the *personality* of the suspect persona; a *question frame* (QF) that is a description of the question of the interviewer; the *interpersonal state* (IS) as 'felt' by the suspect; and an *answer frame* (AF) that holds a description of the answer of the suspect, see Figure 6.2. The suspect's interpretation of the user's question, the *question frame*, influences *the interpersonal state* of the RM, taking into account the *personality* of the suspect and the 'current' interpersonal state. The *answer frame* depends on this (updated) interpersonal state, the question frame, and the personality. The four components of the RM have a resemblance to the 'sense, think, act' cycle with the addition of personality. The question frame holds that which was 'sensed'. The thinking occurs when the interpersonal state and answer frame are updated based on the question frame, interpersonal state, and personality. The acted behaviour of the suspect is based on what is present in the answer frame.

The interpretation of the user's question can be done automatically using social signal processing and natural language understanding techniques. Alternatively, a human wizard can do the interpretation of the user's questions. The wizard can be a tutor or peer in a classroom setting, an experimenter in a lab setting, or the user him or herself. An interpretation need not always be 'as the user intended the behaviour', but interpretations should be consistent. In part II of this thesis, we showed that interpreting social behaviour is not easy. Raters often do not agree on the exact label that a behaviour should have. Compare this to interacting with a human who might not always get what you mean, yet still responds in a consistent manner. We expect this from a virtual human. This means the virtual human does not have to interpret behaviour (and ultimately react) in a way that everyone agrees on. No, the virtual

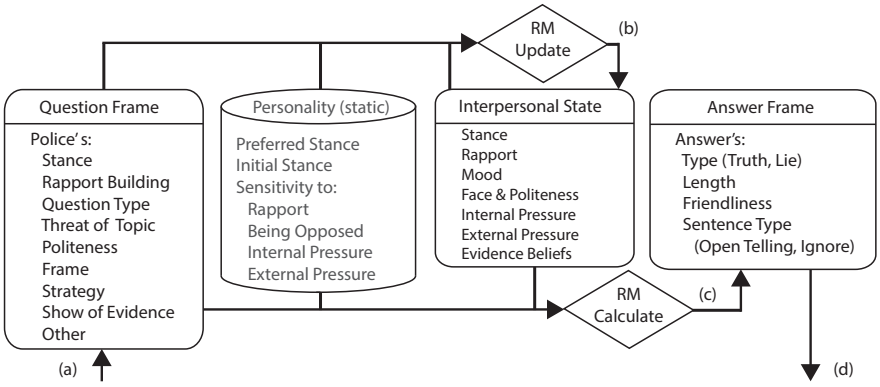

Figure 6.2: A description of the RM. (a) The interrogator asks a question. The values of the question frame (describing the question of the interrogator) are provided to the RM[1]. (b) The RM has a static personality. It has an interpersonal state (which holds values describing how the suspect 'feels') that updates when a question frame is presented. (c) The output of the RM is the answer frame (describing the answer the suspect will give) and it is calculated based on the question frame, personality, and (new) interpersonal state. (d) The answer frame can be used to create or select a behaviour that is consistent with the interpersonal status of the interaction.

[1] *Note that there are several possible perspectives to this description: an egocentric perspective from one observer, for example the point of view of the suspect that is being interrogated. Alternatively an allocentric perspective is an "objective" perspective which, as we saw in chapter 3, can be obtained from a group of raters that do majority voting. We mean the egocentric perspective, and in particular the point of view from the suspect: it is irrelevant whether that is a human-actor, a wizard, or a (collection) of automatic annotation components that do the interpretation of the question for the RM.*

human will have to interpret and react on behaviour of a user in a way that *someone* agrees with.

6.2.1 Variables within the Response Model

Next, we list the variables that are contained within the four components of the RM and discuss each of these variables in more detail.

6.2.1.1 Question Frame

The RM receives as input an interpretation from the user's contribution to the interaction by an interpreter (automatic or manual). The interpreter is not part of the RM. We call this set of input-variables the *Question Frame* (QF). The QF consists of nine aspects that describe how the suspect views the question being posed (see Figure 6.2: Question Frame):

1) The interpersonal stance [116] of the police officer during this contribution, can be: Friendly, Aggressive, Withdrawn, or Dependent. The stances correspond to the four segments of Leary's Rose, not the axes.

2) Rapport building [193] can be done by showing: Attention, Positivity, and Coordination. The amount of rapport the suspect experiences with the user is updated with every contribution of the user.

3) The question type is based on the meta-concept *information* and can be: Open, Yes/No, Probing, Leading, Forced Choice, or Statement. These question types appear frequently in the literature on investigative interviewing (e.g. Snook et al. [177], and Wright & Alison [216]) and were observed in the DPIT corpus.

4) Topic threat describes how face-threatening the topic is for the suspect [28]. This can be: Low, Medium, High, or Guilt Indication[2]. A topic threat that is low, medium, or high does not mean that the topic has to relate to the crime. Topics that relate the suspect to the crime are face-threatening and are coded with Guilt Indication, for example: *"You were seen at the gas station that was robbed yesterday!*".

5) Politeness is related to the politeness strategy used to mitigate a face-threat [28]. Politeness can be coded as: Direct, Approval Oriented, Autonomy Oriented, or Off Record. See section 5.3.2 for a description and examples of the politeness strategies.

6) Dutch police officers go through two phases during an interview: a Person Related Frame that covers the personal life of the suspect and a Case Related Frame that covers topics related to the case [202].

7) Strategy is based on the meta-concept strategy that we observed in police interviews. We operationalize strategy with Giebels' [66] 'Table of Ten' strategies for negotiations. The strategies are: Being Kind, Being Equal, Being Credible, Emotional Appeal, Intimidation, Impose Boundaries, Direct Pressure, Legitimate, Trade, or Rational Convincing.

8) Showing evidence can pressure the suspect into giving up sensitive information or confessing. Using evidence strategically can be very effective for the police in an interview. Summarized, the police can try to convince the suspect that they already know everything by presenting evidence at key moments. This implies that lying will not work and that confessing is the best option, see Granhag and Hartwig [73]. The amount of evidence shown can be: None, Low, or High.

9) The 'Other' attribute is used for special occasions that occur in police interrogations: Confronting a Lie, Repeating the Question, or Accusing.

The variables in the QF are used to describe the properties of the question and social properties of the user at the moment of the question. For example, the user says in a soft voice "*I know it's hard for you to talk about, but it would really help me if you could tell me where you were at the time of the crime*". An interpretation can be: the

[2]It does not matter to the RM whether the suspect is guilty or not. A guilty suspect can be confronted with a topic that is indicative of his guilt. An innocent suspect can be confronted with a topic that implies his guilt.

user is being friendly and understanding, and tries to explain why he needs the information while trying to respect the suspect's autonomy. The open question is highly threatening as it implies that the suspect's whereabouts during the crime is relevant. In the QF this interpretation can be represented as: a `Friendly` stance, building rapport by showing `Positivity` and `Attention`, an `Open` question type, `High` topic threat, `Autonomy Oriented` politeness, a `Case Related Frame`, having a `Being Kind` strategy, showing no new evidence, and without any 'other' attributes.

6.2.1.2 Personality

The *Personality* in the response model consists of a set of static variables of personality traits. The personality influences the calculations that update the interpersonal state (IS) and the answer frame (the AF) of the model. A personality consists of six traits (see Figure 6.2):

1) A *preferred interpersonal stance* that might be considered as a 'personality' and can have the values: Friendly, Aggressive, Withdrawn, or Dependent. It is the stance that the suspect feels most comfortable with, and the stance he or she will try to achieve in conversations. It influences how fast interpersonal stance, mood, and rapport change.

2) The suspect enters the interrogation with a certain stance, his or her *initial stance* in the game. This initial stance is set on the axes dominance and affiliation of Leary's Rose. For example, an aggressive suspect has positive dominance and negative affiliation. This is not a personality trait, however from a game perspective it is important to have an initial value for stance. The variable Initial Stance is located in the personality because it is a static value.

3) The *sensitivity to rapport* states how effective rapport building is with this person. This variable is similar to the Big Five personality trait agreeableness (e.g. [95]). A suspect who scores high on this trait is easygoing, forgiving, and compliant.

4) The *sensitivity to being opposed* of the suspect means how strongly he or she reacts to negative action by the police and how easily he or she turns to aggression. This variable can be compared to the Big Five personality trait neuroticism. A suspect who scores high on this trait is prone to for instance angry hostility and impulsiveness.

5) The *suspect's sensitivity to internal and external pressure* determine whether he will lie or tell the truth when asked about (guilt) sensitive topics. Internal pressure rises with feelings of remorse over one's own actions. For example, a suspect who lied and who is sensitive to internal pressure will experience a strong urge (pressure) to tell the truth.

6) External pressure rises when the police officer puts pressure on the suspect, for example by confronting a lie by showing proof of guilt. The sensitivity to external pressure influences how effective pressure by the police officer is.

It is possible to define a wide range of personalities with these settings. For example, the persona of van Bron from the DPIT-corpus (see Section 3.4.1) can be defined as follows: `preferred stance` aggressive, an `initial stance` of high dominance and low affiliation, low `sensitivity to rapport`, very high `sensitivity to being opposed`, high `sensitivity to internal pressure`, and low `sensitivity to external pressure`.

6.2.1.3 Interpersonal State

The *Interpersonal State* (IS) holds the suspect's idea of the state of the social relation and his or her own 'feelings'. The IS consists of seven variables that are updated during an interaction based on the interpretation of the user's question that is represented in the QF. The variables in the IS are listed here and will be explained next:

1) The suspect's current Stance towards the police officer. The stance is represented by three variables: dominance, affiliation, and a stance label (Leading, Helping, Withdrawn, or Aggressive).

2) The Politeness the suspect experiences in relation to the threat to his or her Face.

3) The current Rapport the suspect experiences with the police officer.

4) The current 'Mood' of the suspect (which can be compliant or aggressive).

5) The Internal Pressure the suspect experiences.

6) The External Pressure the suspect experiences.

7) The number of Evidence Beliefs the police has against him.

1) Suspect's Interpersonal Stance

Interpersonal Stance is "an affective style that spontaneously develops or is strategically employed in the interaction with a person or a group of persons, colouring the interpersonal exchange in that situation" [170, p.705]. 'Leary's Rose' [116], see Figure 1.1, is often used to describe the interplay between dominance and affiliation in an interaction. According to Leary, and others (e.g. [164, 211]), there exist 'interpersonal reflexes' which are relations between the stances of interlocutors. This means that two conversational partners influence each other with their stance during a dialogue. Leary suggested that a together stance invites together stance behaviour in the other and that an opposed stance invites an opposed stance in the other. A dominant stance invites a submissive stance in the other, but a submissive stance invites a dominant stance in the other [116], see Figure 1.1. Police officers in the Netherlands are taught these interpersonal reflexes during their courses on interrogation [202]. They use the theory to try to get an uncooperative suspect in a more cooperative mood, for example by taking a helping stance and hoping the suspect will give in to his "interpersonal reflex" and adopt a cooperative stance. Some suspects are very dominant and some police officers find it challenging to relinquish control of the interrogation to

such a dominant suspect, which can result in a conflict for power. During their training it is emphasized that it can be beneficial for a police officer to take a submissive stance to get the suspect in a cooperative stance.

Scherer noted that "interpersonal stances are often triggered by events, such as encountering a certain person, but they are less shaped by spontaneous appraisal than by affect dispositions, interpersonal attitudes, and, most importantly, strategic intentions" [170, p.705]. In other words, as well as the more or less automatic interpersonal reflexes of Leary there exist personal preferences for a stance and stance taking can be employed strategically. A preference for stance is what Bales [10] called an interpersonal personality. Gurtman [81] suggested that it can be difficult to get someone in a stance that is different from his preferred stance, or what he called the 'predominant interpersonal theme', and people tend to revert back to their preferred stance. In chapter 5, we showed that in police interrogations interpersonal stance is related to face, rapport, 'information exchanges', and strategy.

2) Face & Politeness

Face is a person's public self image and it is something that needs to be defended against acts that threaten it [71]. Brown and Levinson [28] distinguish between negative face, which denote a person's need for freedom (autonomy), and positive face, which denotes a person's need to be approved of and approving of others (approval). Acts by a speaker are potentially face threatening as they might impose on the hearer's face needs. Politeness is a way to mitigate this risk. Brown and Levinson [28] distinguish four types of politeness:

Bald on-record Being straight to the point, e.g., "*Tell me where you were that night.*"

Positive politeness Taking the other's wants into account, e.g., "*Would you like to tell me where you were that night?*"

Negative politeness Not hindering the other's autonomy, e.g., "*If it's not inconvenient to you, could you tell me where you were that night?*"

Off-record Being indirect or vague about one's own wants, e.g., "*I don't seem to have written down where you were that night.*"

A police officer that is bald on record (impolite or direct) has a negative effect on the need for autonomy and approval of the suspect. By employing positive politeness the officer takes the need for approval into account and this has a positive effect on the suspect's need for approval. By taking the other's autonomy into account (negative politeness) the suspect's need for autonomy can be respected. A carefully covered up face threatening act, by being off-record or vague, can respect the suspect's need for autonomy and approval.

3) Rapport

Rapport is the bond that exists between people that have some form of a relationship. Three components of rapport are identified: mutual attention, positivity, and coordination. Rapport and the relative weighting of these components change over

the course of a developing relationship. In early interactions positivity and attentiveness are more important than coordination for the construction of rapport, whereas later coordination and attentiveness are more important [193]. A police officer needs to build rapport with a suspect if he wants to have a productive working relationship. In the DPIT-corpus we saw that police officers attempt to build rapport in early stages of the interrogation by being positive and understanding and asking questions about the suspect's personal life. Later in the interrogation officers ask questions about the case and try to build rapport by listening carefully [31].

4) Mood

Mood is a temporary state of mind or feeling that can be positive or negative, yet 'having a mood' or being moody means having a negative mood. Olsen [139] made a distinction between anger that the suspect directed at the interviewer and anger directed at some outside entity. He stated "Anger at the interviewer will always be considered a negative, whereas anger at the outside person can provide the opportunity for bonding by making the subject feel that the investigator truly understands. The anger at the outside object is the anger in the subject mood. The anger at the investigator is represented by poor rapport" [139, p.4]. A moody suspect is detrimental to the interrogation. Mood was included in the RM given the common occurrence in 'folk psychology' and in the discussions we had with police officers despite that mood has overlaps with interpersonal stance.

5,6,7) Pressure & Proof

Getting a confession from a suspect was long held as the cherry-topping on the results of an interrogation. This high regard for confessions might have had a role in false confessions and convictions as suggested by Kassin [100]. Recently the goal of an interrogation is more information gathering than getting the confession (e.g. [73]), but confessions still occur and for this reason we included the possibility of confessing in the RM. What are the factors that contribute to a confession? Gudjonsson [78] found convicted criminals who confessed gave three reasons for their confessions: internal pressure, external pressure, and proof. The most common reason for confessions is the perception of an overwhelming amount of evidence against the suspect. Internal pressure is defined as those feelings that exert a pressure on the suspect and that the suspect attributes to him or herself, such as feelings of guilt. Correct pressuring of these feelings by the police was rated the second most common reason for criminals to confess. External pressure increases when an external source pressures the suspect, for example repeating questions or confronting lies by presenting proof that refutes the lie. It was found to be most effective with suspect of property offences, younger suspects, and suspects who had never been interviewed by the police before [77, 78].

In the van Bron scenario from the DPIT-corpus, initially the suspect is very aggressive and angry because he has been arrested and because he feels aggression is the only way he can keep some amount of control over his situation. In the RM this initial IS can be defined as follows: `Low Rapport`, `Aggressive Stance`, `Aggressive Mood`, `Low Pressure`, and `Low Evidence Beliefs`.

6.2.1.4 Answer Frame

The response model provides the properties that the response of the suspect should have in the form of an *Answer Frame* (AF) (Figure 6.2). The AF contains four aspects that describe the answer of the suspect:

1) The `AnswerType` is related to the information strategy used by the suspect and can be: Truth, Lie, Avoid, or Aggression.

2) The `AnswerFriendliness` is related to stance and can be: Friendly, Neutral, or Unfriendly.

3) `AnswerLength` is related to the information strategy and can be: Long, Short, One Word, or Silence.

4) `AnswerSentenceType` is related to the question type being posed and the way the suspect wishes to answer to this type and can be: Open Telling, Counter Question, Aggressive Expression, Yes/No, Play Dumb, Probing Answer, or Ignore.

The agent can use the information in the AF and in the IS of the response model to select the most appropriate behaviour in its repertoire. The user asks a question and the system can devise several appropriate answers. The AF can be used to select the most appropriate of these answers. For example, when the police officer asks a very impolite question a suspect is likely to respond in a polite and friendly manner when previously they built rapport and affiliation. However, if they are not 'best buddies' it is likely that the suspect would respond in an aggravated manner to such an impolite question. In this example, the question remains the same and the answers are both providing relevant and requested information but their tone and 'interpersonal meaning' are different: more on how this can be done in chapter 7. The user can respond to this by, for example, asking another question and the cycle of interaction continues.

6.3 Algorithms of the Response Model

In this section we will describe how the RM has been implemented. We will discuss each variable in the RM in turn in the order in which the variables are updated in the RM. We will give a description of the rules and calculations for updating that variable in the RM.

6.3.1 Rapport

In the RM the (integer) variable `Rapport` is located in the Interpersonal State of the virtual suspect. It has a range of 0 to 100. `Rapport` is influenced by `Rapport Building`, `Strategy`, and `Frame`, see Table 6.1. `RapportBuilding` consists of three Booleans: positivity, attention, and coordination. `Frame` is a variable that indicates the phase of the interrogation (early or later). `Strategy` is a variable that indicates the conversational strategy that the officer applies, based on the 'Table of

Table 6.1: For Rapport the RM uses the variables under IN to calculate the variables under OUT.

Rapport			
IN		OUT	
RapportBuilding	(QF)	Rapport	(IS)
Strategy	(QF)		
Frame	(QF)		

Table 6.2: For Face the RM uses the variables under IN to calculate the variables under OUT.

Rapport			
IN		OUT	
Politeness	(QF)	FaceOfSuspect	(IS)
		- Autonomy	
		- Approval	

Ten' by Giebels [66]: for example being kind or intimidating. Rapport increases when there is RapportBuilding. This increase is largest when, during the early stages of the interrogation (indicated by Frame), there is attention and positivity or when during later stages of the interrogation there is attention and coordination. Rapport decreases when there is no RapportBuilding or when the officer uses the Strategy 'Intimidation'. To our knowledge there is no information available in the literature as to what the size of the increase and the decrease of rapport should be and so the sizes have to be guessed for the RM.

6.3.2 Face

In the RM the face of the suspect (FaceOfSuspect) is dependent on the type of Politeness employed by the officer, see Table 6.2. A police officer that is bald on record (impolite or direct) has a negative effect on the need for autonomy and approval of the suspect. By employing positive politeness the officer takes the need for approval into account and this has a positive effect on the suspect's need for approval. By taking the other's autonomy into account (negative politeness) the suspect's need for autonomy can be respected. A carefully covered up face threatening act, by off-record or vague, can respect the suspect's need for face and autonomy, see Table 6.3. The suspect's face-needs are used to calculate the interpersonal stance of the suspect.

6.3.3 Interpersonal Stance

In the RM the suspect's stance is represented by three variables: X indicating the affiliation the suspect has with the police officer, Y indicating the dominance of the suspect, and a 'CurrentStance'. The X and Y values of stance have a range of -100 to 100. The X,Y-coordinates indicate in which current stance the suspect is on Leary's Rose, see Figure 1.1. The suspect's interpersonal stance is influenced by the (previous) stance of the suspect (SuspectStance), the stance of the officer (PoliceStance),

Table 6.3: The effect of the politeness strategy used by the police officer on the face-needs of the suspect.

	Suspect's Face	
Politeness	Autonomy	Approval
Bald on-record	-1	-1
Positive politeness	0	1
Negative politeness	1	0
Off-record	1	1

the suspect's personality, `RapportBuilding` by the officer, `Strategy` of the officer, the threat of the questions (`TopicThreat`) and the `Politeness` employed by the officer. The effects of these variables can vary depending on the phase (`Frame`) of the interrogation, see Table 6.4.

Table 6.4: For Interpersonal Stance the RM uses the variables under IN to calculate the variables under OUT.

Interpersonal Stance			
IN		OUT	
`PoliceStance`	(QF)	`SuspectStance`	(IS)
`Frame`	(QF)	- `X`	
`RapportBuilding`	(QF)	- `Y`	
`Strategy`	(QF)	- `CurrentStance`	
`TopicThreat`	(QF)		
`SuspectStance`	(IS)		
`FaceOfSuspect`	(IS)		
`PreferredStance`	(Pers)		
`Sens.to`	(Pers)		
- `Opposition`			
- `Int.Press.`			
- `Rapport`			

6.3.3.1 Stance X: Affiliation

The X variable of the stance of the suspect, his affiliation with the police officer, is calculated (updated) by the RM with the equation:

$$\begin{aligned} Susp.StanceX = Susp.StanceX &+ \frac{Pref.StanceX}{Pref.StanceXInfl.} \\ &+ \frac{Pol.StanceX}{Pol.StanceXInfl.} + MirroredStance \\ &+ RapportEffect + ApprovalEffect \\ &+ StrategyEffect \end{aligned} \tag{6.1}$$

The stance of the suspect is his 'old stance' plus the influences on his stance. The preferred stance is a constant from the personality of the suspect. The suspect's stance slowly moves in the direction of the suspect's preferred stance [81]. The stanceX of the police officer indicates how affiliated the officer behaves towards the suspect. The suspect's stance will move towards the stance of the police officer. PreferredStanceX-Influence and PoliceStanceXInfluence are constants that influence the size of the effect of the preferred stance and the stance of the officer (these sizes were guessed as they were not available in the literature).

People tend to favour a stance that is similar to their preferred stance [81]. In the RM, the stanceX (affiliation) of the suspect is influenced by the stanceY (dominance) of the officer. A suspect 'likes' it if an officer has a stanceY that is opposite of the preferred stanceY of the suspect and 'dislikes' it if they have the same stanceY. *MirroredStance* expresses this liking or disliking by increasing or decreasing the stanceX of the suspect respectively.

The effects of rapport, approval (face), and strategy on the suspect's stanceX are calculated separately. Rapport building by the officer has a positive effect (*RapportEffect*) on the stanceX of the suspect. This effect is larger during the early phase of the interrogation (person frame).

The effect the face-need for approval of the suspect (*ApprovalEffect*) has on his stance depends on the threat of the topic discussed and the frame of the interrogation, see Table 6.5. The stanceX of the suspect increases when the face-need for approval of the suspect is met, but decreases when the face-need is violated. In the early phase of the interrogation (Person Frame) the effect of approval is lower when the threat of the topic discussed is higher. This means for example, a suspect likes a police officer best when they discuss topics that are not face-threatening and the officer acknowledges the suspect's face-needs. Later in the interrogation (during the Case Frame) the topic is the case. The suspect expects questions about the case and thus is better prepared for threatening questions. The negative effect of threatening topics is smaller than in the early stages of the interrogation, see Table 6.5.

The conversational strategy employed by the officer has an effect (*StrategyEffect*) on the stanceX (affiliation) of the suspect [66]. Using intimidation as a conversational strategy has a negative effect on suspects who have sensitivity towards opposition, meaning they will move towards an opposed stance. Conversational strategies 'being kind' or 'being equal' has a positive effect on stanceY. This effect is larger for suspects who have a personality with a higher preferred stanceY (affiliation). Emotional appeals by the officer increase the stanceX of the suspect. Suspects who are sensitive to internal pressure respond stronger to emotional appeals.

6.3.3.2 Stance Y: Dominance

The Y variable of the stance of the suspect, his dominance in relation to the police officer, is calculated (updated) by the RM with the equation:

$$Susp.StanceY = Susp.StanceY + \frac{Pref.StanceY}{Pref.StanceYInfl.} - \frac{Pol.StanceY}{Pol.StanceYInfl.} + AutonomyEffect \tag{6.2}$$

Table 6.5: The effect the face-need for approval has on stance (ApprovalEffect) depends on the threat of the topic discussed and the frame of the interrogation.

	Person Frame			**Case Frame**		
Approval → **TopicThreat ↓**	-1	0	1	-1	0	1
Low	-5	0	10	0	0	1
Medium	-10	0	8	-5	0	4
High	-20	0	5	-7	0	6
GuiltIndication	-20	0	-10	0	0	10

Table 6.6: The effect the face-need for autonomy has on stance depends on the threat of the topic discussed and the frame of the interrogation.

	Person Frame			**Case Frame**		
Autonomy→ **TopicThreat↓**	-1	0	1	-1	0	1
Low	5	0	0	5	0	0
Medium	8	0	0	7	0	0
High	12	0	0	10	0	0
GuiltIndication	15	0	0	10	0	0

The stance of the suspect is his 'old stance' plus the influences on his stance. The preferred stance is a constant from the personality of the suspect. The suspect's stance slowly moves in the direction of the suspect's preferred stance [81]. The stanceY of the police officer indicates how dominantly the officer behaves towards the suspect. The suspect's stance will move away from the stance of the police officer. This means that if an officer behaves dominantly the suspect will move to a more submissive stance and if the officer behaves submissively the suspect will move to a stance that is more dominant. PreferredStanceXInfluence and PoliceStanceXInfluence are constants that influence the size of the effect of the preferred stance and the stance of the officer (these sizes were guessed as they were not available in the literature).

The effect the face-need for autonomy of the suspect (*AutonomyEffect*) has on his stance depends on the threat of the topic discussed and the frame of the interrogation, see Table 6.6. The stanceY (dominance) of the suspect increases when the face-need for approval of the suspect is violated. This increase is larger when the threat of the topic is higher. This means that the suspect becomes more dominant when the officer is impolite and addresses a topic that is threatening to the suspect.

Table 6.7: The rules that determine the positive and negative influences on the suspect's mood, see main text for details. An asterisk (*) indicates a formula and formulas are displayed below the Table.

		Influence	
Rule		**Pos**	**Neg**
#1	Previous Mood:		
	- (Aggressive)		40
	- (Compliant)	20	
#2	Strategy:		
a	- (Intimidation)		25
	& PrefStance(Aggr)		50
b	- (Being Kind)	15	
	& PrefStance(!Aggr)	30	
c	- (Being Equal)	15	
	& PrefStance(!Aggr)	30	
d	- (Emotional Appeal)	*[1]	
	& SuspectStance($Y < 0$)	50	
e	- (Direct Pressure)		
	& Frame(Person)		
	& PrevAnsw(Avoid || Aggr)		
	& RepeatQ(true)		25
f	- (Direct Pressure)		
	& Frame(Case)		
	& PrevAnswer(Lie)	25	
#3	SuspectStance:		
	- (Aggressive)		50
	- ($X > 0$)	50	
#4	Sensitivity to Opposition		*[2]

[1] $(Pos) = (SensToIntPress) * 40/100$
[2] $(Neg) = (Neg) * (SensToOpposition)/100$

6.3.4 Mood

The `Mood` of the suspect in the RM can be 'compliant' or 'aggressive'. The variables from Table 6.8 under IN can have values that have a positive or negative influence on the suspect's mood. The total positive and negative influences are calculated following the rules in Table 6.7. The largest influence, positive or negative, 'wins'. The (sequential) rules in the RM to determine the mood of the suspect are defined as follows (see Table 6.7 for the size of the influences and the logic):

Table 6.8: For Mood the RM uses the variables under IN to calculate the variables under OUT.

Mood			
IN		OUT	
Frame	(QF)	Mood	(IS)
Strategy	(QF)		
TopicThreat	(QF)		
Other - RepeatQuestion	(QF)		
SuspectStance	(IS)		
Mood	(IS)		
Pref.Stance	(Pers)		
Sens.to - Opposition	(Pers)		
- Int.Press. AnswerType	(AF)		

#1 The previous Mood of the suspect influences the new mood. The mood of a person is more likely to be bad when suspected of a crime and in an interrogation. If the previous mood was negative it is likely to remain bad, if the mood was good it might stay good.

#2 The conversational Strategy [66] used by the officer can have an influence on the mood of the suspect.

a An officer that uses intimidation has a negative influence on the mood of the suspect. This negative effect is larger when the suspect has an aggressive personality (PreferredStance).

b An officer that is being kind to a suspect will improve his mood, this is especially true for suspects who do not have an aggressive personality.

c An officer that treats a suspect like an equal will improve the mood of the suspect and this effect is larger when the suspect does not have an aggressive personality.

d An emotional appeal is aimed at the Internal Pressure of the suspect, for example 'do what is right'. Feeling high internal pressure makes a suspect more likely to cooperate which is represented here as a positive effect on mood. The effect of an emotional appeal is stronger for suspects that are Sensitive to Internal Pressure, see formula 1 under Table 6.7. In addition, an emotional appeal is more effective on someone who has a submissive Stance.

e Direct pressure is defined as "in a neutral fashion apply pressure on the other by being steadfast" [66, p.149]. In police interrogations [31] we observed direct pressure to have a negative effect when the interrogation was in the 'person'-phase. In this phase the officer tries to get to know the suspect and asks predominantly personal questions. Applying direct pressure in this Frame seems

rude, especially when the suspect previously avoided answering the personal question. This has a negative effect on the mood of the suspect.

f In the phase where the case is discussed, direct pressure is often applied, for example, to confront a lie in a `Previous Answer`. This can lead to an admission or more cooperation which is represented in the RM as a positive effect on mood.

#3 The `Stance` of the suspect has an effect on his mood. An aggressive suspect is more likely to be in a bad mood, whereas a suspect who has high affiliation is likely to get a positive mood.

#4 A suspect who has low `Sensitivity to Opposition` is likely to remain composed when he experiences 'negative things'. In the RM the sensitivity to opposition can range from 0 to 100. The 'negative'-score is divided by the fraction of this sensitivity.

The largest influence, positive or negative, 'wins'. The mood of the suspect is set accordingly: if positive is larger than negative the mood of the suspect becomes 'compliant' otherwise it is set to 'aggressive'.

6.3.5 Internal Pressure

In the RM, the variables in Table 6.9 under IN have an influence on the `Internal Pressure` of the suspect. Internal pressure is an integer in the range 0 to 100. If a topic that is indicative of the guilt of the suspect (i.e. a high `TopicThreat`) is addressed, the internal pressure increases. This increase is larger when the officer also uses an emotional appeal (`Strategy`). In this case, the officer plays on the emotions of the suspect to increase feelings of guilt. The increase in internal pressure is also larger when the officer confronts the suspect with a lie, as this makes the suspect feel guilty about lying. Submissive personalities are more perceptible to feelings of guilt (e.g. [79]) and thus when internal pressure increases the size of increase is larger for submissive suspects. The size of the increase depends on the `Sensitivity to Internal Pressure` of the suspect. The internal pressure decreases when the suspect tells the truth (see Section 6.3.8).

6.3.6 External Pressure

In the RM, the variables in Table 6.10 under IN have an influence on the `External Pressure` of the suspect. External pressure is an integer in the range 0 to 100. If the `TopicThreat` is high and indicating guilt, the external pressure increases. This increase is larger when there is also a lie that is confronted or a question is repeated. Also when the interrogator uses an intimidating or direct `Strategy` external pressure of the suspect increases. The size of the increase of external pressure depends on the suspect's `Sensitivity to External Pressure` and the suspect's `PreferredStanceY`. The increases are larger for suspects with high sensitivity for external pressure and for submissive personalities (`PreferredStanceY` < 0). When the suspect tells the truth the external pressure decreases.

Table 6.9: For Internal Pressure the RM uses the variables under IN to calculate the variables under OUT.

Internal Pressure			
IN		OUT	
TopicThreat	(QF)	Int.Pressure	(IS)
Strategy	(QF)		
Other	(QF)		
- LieConfronted			
Int.Pressure	(IS)		
Rapport	(IS)		
Pref.StanceY	(Pers)		
Sens.to	(Pers)		
- Int.Pressure			
- Rapport			

Table 6.10: For External Pressure the RM uses the variables under IN to calculate the variables under OUT.

External Pressure			
IN		OUT	
Strategy	(QF)	Ext.Pressure	(IS)
TopicThreat	(QF)		
Other	(QF)		
- LieConfronted			
- RepeatedQuestion			
ExternalPressure	(IS)		
Pref.StanceY	(Pers)		

6.3.7 Evidence Beliefs

A suspect has an idea about the amount of evidence the police officer has about the crime he committed. This belief does not necessarily correspond with the amount of evidence the officer really has. By using his evidence tactically, the officer can make the suspect believe he holds more evidence than he does [73].

In the RM the `Evidence Beliefs` is represented by an integer (ranged 0-100) and is influenced by the variables in Table 6.11. The evidence belief of the suspect increases when the police officer presents (new) evidence. Additionally, when the suspect admits to a crime, he tells the truth (which is represented in the RM in the `AnswerType`), his belief in the evidence held by the interrogator will increase.

6.3.8 Answer Type

In the DPIT corpus we observed four categories of answers that suspects often gave: Truth, Lie, Avoid, or Aggression. The most obvious distinction in the answers a suspect

Table 6.11: For Evidence Beliefs the RM uses the variables under IN to calculate the variables under OUT.

Evidence Beliefs			
IN		OUT	
`ShowOfEvidence`	(QF)	`EvidenceBeliefs`	(IS)
`EvidenceBeliefs`	(IS)		
`AnswerType`	(AF)		

can give is whether he or she is lying or telling the truth. Another option available to a suspect is to avoid answering the question altogether. Avoiding is fundamentally different from lying or telling the truth, it is withholding information that was requested by providing information that is (somewhat) unrelated to the requested information. This provided information might be another lie or the truth. Blume and Board [25] call this 'vagueness' and argue this is often used intentionally. An avoiding answer type can take a variety of forms that can be strategically employed. Avoiding the answer can delay giving the requested information or it might steer the interlocutor to a different topic. For example, the police officer asks "*What work do you do?*" to which the suspect gives the avoiding answer "*What do you want to know? I mean... what work do you do?*". A suspect can utilise aggression to avoid giving an answer. In the DPIT corpus this often involved questioning the question and the person that asked the question. Aggression occurred most with suspects that had an aggressive personality (e.g. van Bron, see Section 3.4.1) and when the police officer talked about topics that the suspect perceived as threatening: very personal questions and topics that were related to the crime. For example, a police officer asked an aggressive unemployed suspect "*What work do you do?*" to which the suspect responded with in aggressive tone "*Why are you asking me about my work? Come on! What is this nonsense?! Are we gonna talk about my work?*".

In the RM the `AnswerType` is determined based on the variables under IN in Table 6.12. A suspect in an aggressive `Mood` will give an aggressive answer. In other cases the `AnswerType` depends on the `Frame` of the interrogation. In the 'person related' frame the police officer asks questions about the personal life of the suspect. In the person-frame different topics are perceived as threatening, for example questions about the suspect's car might not be perceived as threatening. Whereas questions about the suspect's car are perceived as very threatening when the car was involved in the crime. This has implications for the `AnswerType` of the suspect.

In the person related `Frame` the suspect will tell the truth if the interpersonal relation is sufficiently positive. This is also true when the topic is threatening, but a higher level of `TopicThreat` requires that the interpersonal relation needs to be higher for the suspect to tell the truth. Here, positive interpersonal relation means that the suspect experiences high `Rapport`, positive affiliation (`SuspectStanceX`), and has a positive `Mood`. A lower interpersonal relation or a higher topic threat means the AnswerType will deteriorate to avoid, lie, or even aggression. We observed in the DPIT corpus that a question or statement that indicated the suspect was guilty was

often countered with aggression. Thus in the RM, when the topic is indicative for the guilt of the suspect, the AnswerType is aggressive. The exception being a suspect that likes the interrogator (SuspectStanceX> 0) and does not have an aggressive PreferredStance.

In the case related Frame the suspect is confronted with questions about the crime. During case related questions the suspect will tell the truth if the interpersonal relationship is sufficiently positive. When a topic is more threatening the interpersonal relationship has to be more positive before a suspect tells the truth, which is similar to the person related Frame. If the interpersonal relationship is not positive enough, the suspect will avoid, lie, or even turn aggressive. In a police interrogation the officer will try to 'surround a fact' with statements from the suspect, see chapter 1. This leads to an eventual confrontation where the officer confronts the suspect with inconsistencies in the story told. These confrontations increase the pressure the suspect experiences. Whether a suspect tells the truth depends, among other things, on the amount of evidence he or she believes the police has and on the pressure the suspect feels. Default, the suspect will try to avoid giving information that is incriminating and thus lie about the details of his or her crime[3]. A suspect wants to appear credible (e.g. [73]) and thus will admit to a crime when the evidence is overwhelming. In the RM, confrontations and accusations are always on threatening topics as they are indicative for the guilt of the suspect. A suspect whose EvidenceBeliefs are very high (max $= 100$) will tell the truth when he or she faces guilt indicating questions during the case related Frame, even when he or she is facing an Accusation. Some additional convincing is required when the suspect's EvidenceBeliefs are high (> 70). For example, the suspect will tell the truth if the officer applies pressure, by raising the internal and external pressure, or uses the Strategy "rational convincing" [66]. A suspect that has an aggressive Stance will lie or avoid the question or turn to aggression, even when his evidence beliefs are high. In all other cases when the topic is guilt indicating, the response of the suspect is to try to lie or avoid the question.

Based on the AnswerType several variables in the IS can change. Internal and External Pressure are reduced when the suspect gives a truthful AnswerType. When the suspect tells the truth on a topic that is related to the case, the EvidenceBeliefs will increase, as the police officer now holds more evidence according to the suspect. The affiliation, SuspectStanceX, is reduced if the suspect turns to an aggressive answer meaning the suspect will 'like' the interrogator less.

6.3.9 Answer Sentence Type

The Sentence Type of the answer of the suspect describes the form of the answer: the type of sentence used as a response. These responses are based on the question types

[3]For AnswerType it makes little difference in the RM whether the suspect is guilty or not. An innocent suspect who tells the truth during a confrontation will tell he or she is innocent of the crime. A 'confession' in this case can mean the suspect does reveal new (to the police) information. A guilty suspect can confess to the crime when telling the truth. The RM does not contain the scenario of the suspect and the facts that can be discussed. It contains not "what" should be said, only the "how" it should be said, see Figure 6.1.

Table 6.12: For the AnswerType the RM uses the variables under IN to calculate the variables under OUT.

AnswerType			
IN		OUT	
TopicThreat	(QF)	AnswerType	(AF)
Frame	(QF)	SuspectStanceX	(IS)
Strategy	(QF)	Int.Pressure	(IS)
Other	(QF)	Ext.Pressure	(IS)
- Accusation		EvidenceBeliefs	(IS)
SuspectStance	(IS)		
Rapport	(IS)		
Mood	(IS)		
Int.Pressure	(IS)		
Ext.Pressure	(IS)		
EvidenceBeliefs	(IS)		
Pref.Stance	(Pers)		
Sens.to	(Pers)		
- Rapport			

Snook et al. [177] describe and that frequently occur in the literature on interrogation (e.g. [216]). We distinguish the following Answer Sentence Types:

- **Open telling:** An open telling account from the suspect revealing much of the information the suspect has. For example, to the question "*What did you do?*" an open telling answer can be "*I went to the market, bought some apples, went home, did some laundry, watched some TV, cooked a meal and went to work after that. Then came home from the night shift and went straight to bed.*".

- **Counter question:** A response in the form of a question. This occurs most with suspects that are trying to avoid giving an answer, to buy time or contest the question. For example to the question "*Are you married?*" a counter question could be "*Are YOU married?*".

- **Aggressive expression:** When suspect answers with an aggressive `AnswerType`, it is reflected in the sentence. For example, an aggressive response to any question can be "*Piss off, copper!*".

- **Yes/No:** The yes/no answer sentence type is appropriate when answering a closed question. However, it can reveal more information when the suspect wants to give a longer answer. For example, to the question "*Were you at the gas station?*" a longer yes/no answer can be "*Yes, I was at the gas station.*".

- **Play dumb:** An appropriate answer when the suspect plays dumb or genuinely does not know, for example "*I don't know.*".

Table 6.13: For the Answer Sentence Type the RM uses the variables under IN to calculate the variable under OUT.

AnswerSentenceType			
IN		OUT	
`QuestionType`	(QF)	`AnswerSentenceType`	(AF)
`AnswerType`	(AF)		

- **Probing answer:** The probing answer sentence type is appropriate when answering about a single topic in a compound question. For example to the question "*Who else was there, were there others, could you see the whole car park?*" to which the suspect can answer "*There was no one else.*".

- **Choice:** The suspect picks one of the choices the interrogator presents. For example, when faced with the choice "*Was the car red or blue?*" the suspect can choose "*Red.*".

- **Ignore:** Ignoring the interrogator's question altogether. This can be silence but might also be a response that is unrelated to the question. For example ignoring the question and pointing out the window "*Hey, a bird!*".

In the RM, the `AnswerSentenceType` is determined based on the variables under IN in Table 6.13. It is related to the type of question that the interrogator posed and the type of answer the suspect wishes to give. Table 6.14 shows this relation between `QuestionTypes`, `AnswerTypes`, and `AnswerSentenceTypes`.

6.3.10 Answer Length

Longer is better when it comes to the length of the answer that a suspect gives. The police attempt to gather information from the suspect in an interrogation and a longer answer provides more information. The length of an answer is influenced by whether or not a suspect is lying. Analysis of interrogations of (mock) crimes by Strömwall et al. [183] indicated that liars used simpler answers with less detail than truth-tellers who 'told it as it was'. In addition, Richardson et al. [160] showed that interviews that lead to a confession have a higher rate of the suspect matching the verbal language style of the interviewer than interviews that did not lead to a confession. Language style matching might be considered as a form of rapport in the form of coordination [193]. We take this to mean that feelings of affiliation or rapport will lead to more revealing and truthful answers.

In the RM, `AnswerLength` is determined based on the variables under IN in Table 6.15. The `AnswerLength` can be: Long, Short, One Word, or Silence. Table 6.16 shows the effect of `AnswerType`, `AnswerSentenceType`, and whether or not the suspect experiences positive affiliation or rapport on the `AnswerLength`.

Table 6.14: The options for the Answer Sentence Type, based on the Question Type and the Answer Type.

QuestionType	AnswerType	AnswerSentenceType
Open	Truth / Lie	Open Telling
	Avoid	Counter question
	Aggressive	Aggressive expression
Yes/No	Truth / Lie	Yes/No
	Avoid	Play dumb
	Aggressive	Aggressive expression
Probing	Truth / Lie	Probing answer
	Avoid	Play dumb
	Aggressive	Aggressive expression
Leading	Truth / Lie	Yes/No
	Avoid	Play dumb
	Aggressive	Aggressive expression
Forced choice	Truth / Lie	Choice
	Avoid	Play dumb
	Aggressive	Aggressive expression
Statement	Truth / Lie	Open Telling
	Avoid	Ignore
	Aggressive	Aggressive expression

Table 6.15: For the Answer Length the RM uses the variables under IN to calculate the variable under OUT.

AnswerLength			
IN		OUT	
SuspectStance	(IS)	AnswerLength	(AF)
Rapport	(IS)		
Sens.to - Rapport	(Pers)		
AnswerSentenceType	(AF)		
AnswerType	(AF)		

6.3.11 Answer Friendliness

In the RM, AnswerFriendliness is determined based on the variables under IN in Table 6.17. The friendliness of the answer of the suspect can be: Friendly, Neutral, or Unfriendly. If the Mood of the suspect is aggressive, the AnswerFriendliness will always be unfriendly. Questions with a high TopicThreat require a more positive stance and higher rapport to be answered in a friendly manner compared to topics that are less threatening. Topics that are indicative of guilt are answered in an un-

Table 6.16: The length of the answer depends on the AnswerType, the AnswerSentenceType and whether or not the suspect experiences affiliation, rapport, or both with the interrogator.

AnswerType	Answ.Sent.Type	Affiliation or Rapport	AnswerLength
Truth	Open telling	+	Long
		-	Short
	Yes/No	+	Short
		-	One word
	Probing answer	+	Short
		-	One word
	Choice	+	Short
		-	One word
Lie	Open telling		Short
	Yes/No		One word
	Probing answer		One word
	Choice		One word
Avoid	Counter question		Short
	Play dumb		Short
	Ignore		Silence
Aggressive	Aggressive		Short

friendly, or neutral manner when there is sufficient rapport and a positive stance. Suspects with an aggressive personality (PreferredStance) answer in a less friendly manner than other suspects.

Table 6.17: For the Answer Friendliness the RM uses the variables under IN to calculate the variable under OUT.

AnswerFriendliness			
IN		OUT	
TopicThreat	(QF)	AnswerFriendliness	(AF)
SuspectStance	(IS)		
Mood	(IS)		
Rapport	(IS)		
PreferredStance	(Pers)		
Sens.to - Rapport	(Pers)		
AnswerType	(AF)		

6.4 Using the Response Model

In this section I will describe how to use the RM and what other components are needed for a virtual suspect that can understand the user and respond appropriately. Currently, the RM is implemented in JAVA.

6.4.1 Interfacing with the Response Model

The RM deals with the 'how-things-are-said' of a conversation and as such is not meant as a standalone application[4]. The RM is envisioned for use in an application that features speech recognition, reasoning about what to say, and an embodied virtual character that plays the role of the suspect: a 'virtual suspect' system.

In order to use the RM in a broader application, we created an interface. Three actions are possible in this interface:

1) The RM needs to be initialised with a personality. This needs to be done once before the interaction starts.

2) During the interaction, at certain intervals or when there is new information available, the RM can be sent a Question Frame (QF). Whenever the RM receives a QF it updates the Interpersonal State (IS) and the Answer Frame (AF). The RM requires that the QF is complete, that means that all the variables in the QF have a valid value.

3) The IS and the AF are available for request through the interface at every moment after the initialisation and not during an update of the IS and AF. Whenever requested, the RM provides the current state of the IS and AF. These can then be used by other components in the virtual suspect system to show appropriate behaviour.

One of the interfaces with the RM we developed offers a GUI for a Wizard of Oz setup, see Figure 6.3. Here a human observer plays the role of interpreter for the RM providing question frames.

6.4.2 Updates

The best moment for requesting an update of the RM, that is presenting a QF, is difficult to determine. As Chindamo et al. [44] suggested and we found again in chapter 3, it is difficult to obtain a reliable and meaningful annotation of fuzzy social aspects such as stance on a short timescale. Chindamo et al. [44] suggested, for interpersonal stance, annotations on about 15 minute segments. This is longer than a typical interaction with a virtual human lasts. Therefore, we suggest that in a Wizard of Oz setup the wizard takes the lead and determines when he or she feels a meaningful (new) QF can be provided. As a rule of thumb we suggest a new QF be selected and the RM is updated roughly every minute or whenever an apparent

[4]Although it might be an educational experience for some users to interact with the RM using question and answer frames. Such an abstract interaction can make a user reflect on his interaction style in an abstract manner. This reflection can be a valuable learning experience [105].

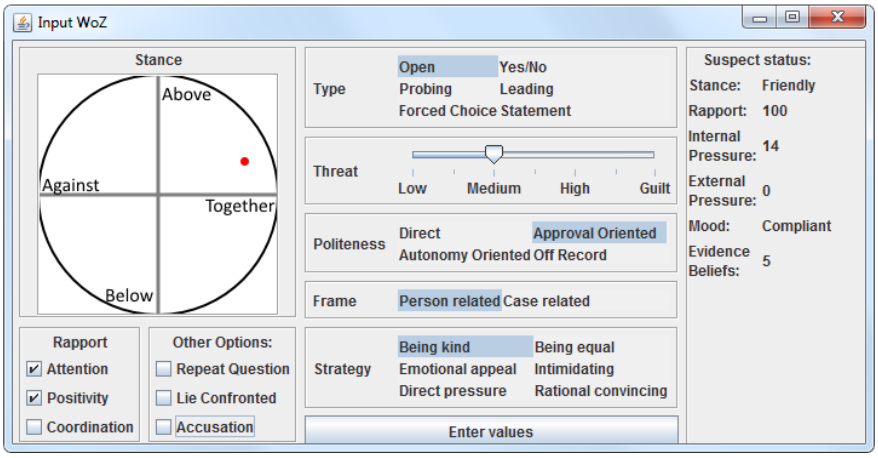

Figure 6.3: A GUI for a Wizard of Oz setup. The wizard can provide a (simplified) question frame to the RM. The (simplified) status of the suspect is shown on the right.

change in the interaction style is perceived by the wizard. A complete QF has to be provided to the RM in the current implementation. This means when one variable in the QF might have changed all variables have to be re-sent and thus will be used to update the RM state. It is questionable whether these unchanged variables contribute meaning to the update: they have again an influence. In an automated annotation or interpretation of the user's behaviour, some modules might be able to provide an update to the QF very often, giving the possibility of a runaway update process.

6.5 Conclusions

In this chapter I presented our RM for a virtual suspect based on the literature and on information obtained from the DPIT-corpus. We need to know whether the RM can portray a persona in a recognizable and consistent way. In the next chapters I will discuss some of the evaluations we conducted of the RM.

6.5.1 Limitations and Future Work

The RM presented in this chapter describes the effects that interpersonal relationship has on verbal contributions of a virtual suspect. It does not yet explicitly model the non-verbal behaviour of a virtual suspect. It is clear that the non-verbal behaviour is very important in police interrogations. Some of the variables in the RM can be utilized for the selection of appropriate non-verbal behaviours, for example the 'typical stance behaviours' discussed in chapter 4 could be matched to the suspect's interpersonal stance during the interrogation. Finally, partially updating the RM or updating the RM on a partial QF is not possible in the current implementation. This limitation should be addressed in future work.

7

A Method to Evaluate Response Models

In this chapter I will report on our efforts to evaluate the response model I described in the previous chapter. This is an algorithm that computes the responses of an agent that plays the role of a suspect in simulations of police interrogations. The response model is centred around keeping track of interpersonal relations. The model is parametrized in such a way that different personalities of the virtual suspect can be defined. In the evaluation we defined three different personalities. We had participants guess the personality based on the responses the model provided in an interaction with the participant: the 'Guess who you are talking to?' test. I will describe two experiments which vary in the manner participants interacted with the response model. In the first experiment participants had an abstract interaction, a meta-interaction in which participant and response model exchanged a series of abstract descriptions of 'how-I-would-say-it'. In the second experiment participants interacted with an embodied virtual suspect that could understand what they said and how they said it. We will investigate what factors contributed to the ability of our virtual agent to show behaviour that was recognized by participants as belonging to a persona.[a]

[a] This chapter is based on two publications [32, 34]

7.1 Introduction

Credible virtual human responses are crucial for a serious gaming system that will train police officers to do a proper interrogation. If players are to engage in a meaningful social interaction with an artificial agent, they first need to suspend their disbelief in the realism of the artificial interaction (e.g. [138]). Ströfer et al. [181] interrogated participants about a transgression they committed using an avatar and

while recording their electrodermal activity (EDA). Truth tellers and deceivers could be distinguished based on their EDA only when the participant believed the avatar was controlled by a human. They conclude that "the belief that one is talking to a human instead of an autonomous computer-operated system" is crucial for a real-life application [181]. In gaming this suspense of disbelief is aided by creating a compelling narrative, realistic virtual environments, events, non-player characters, and the consistency between these. Perlin [146, p146] put it like this, "characters should be free to respond in a way that accords with their nature". We focussed on creating a system in which actions of the user will be met with credible responses of the virtual human: a response model. As Ochs et al. put it "Credibility, [...], relies greatly on the notion of consistency " [138, p281]. They go on to distinguish two dimensions of consistency in artificial characters (in games):

1 'Consistency with past behaviour' which entails that the behaviour of an artificial character should match its personality and the events that preceded a 'current' behaviour.

2 'Consistency with the current environment'.

Our efforts towards this consistency took the form of a response model (RM) that can be set to a personality and takes into account the interpersonal events that have occurred in the interaction in an interpersonal state that is updated with new events. This state is used to compute an answer frame that abstractly describes the response of the artificial agent, see chapter 6. In this chapter we will present our experiment to investigate whether users can differentiate between different personality-settings in the RM and whether they can agree on a description of the virtual suspect. We investigatde using a 'Guess who you were talking to'-test. In this test, participants interacted with the response model and had to guess which persona was portrayed by the system.

We will also present a way to evaluate *only* the response model and not any other component in a virtual human system. We did this by having participants interact with the RM without linguistic content. This is an abstract interaction in terms of the RM: the user asks 'questions' in the terms of the *Question Frame* and the RM presents its 'answers' in terms of the *Answer Frame*. The frame abstracts from the semantic context, for example an officer makes a *friendly statement* to which the virtual suspect responds with a *friendly question for clarification*:

Police "*We would like your help to solve the mystery of the missing cookies.*"

Suspect "*Sure, I'll do what I can, but which cookies do you mean? The chocolate or the vanilla?*"

First we will evaluate the RM in an experiment where participants engage in such an abstract interaction with the RM without linguistic content. After that we will describe an experiment where participants used natural language to interact with the virtual suspect. In both, participants played the role of police interviewer and the RM played the role of suspect. By comparing the two experiments we can gain insight into the source of recognition errors.

7.2 Method for Evaluation of Response Models

The problem with evaluating a virtual human is that it often remains unclear what each component in the system contributes to the outcome of the evaluation. For example, the cause of inappropriate behaviour might be in the virtual human's speech recognition, interpersonal or emotional interpretation, reasoning, or the authoring of the response behaviours available to the system. We suggest an evaluation of the RM in which participants interacted with the RM using the terms from the *question frame* and the *answer frame* without having to formulate their question in natural language. Evaluating a response model in a system that uses natural language messages, which have a subjective quality, introduces (at least) two extra sources of ambiguity:

1) The user says something which has to be interpreted by the system or a Wizard of Oz. As we showed in chapter 3 it is hard to get reliable interpretations of subjective measures such as interpersonal stance, even for trained annotators. It is likely that such interpretations will introduce errors into the system, where an error is that the interpretation is different from the intended message. For example, the user says something that is ambiguous, it might be interpreted as friendly banter or an aggressive expression, and the interpretation is different from the way it was intended.

2) The author of the scenario, utterances, and behaviours available to the virtual human has an intention with the virtual human's behaviours. This intention can be different from that which the user thinks these behaviours meant, giving rise to confusion. In other words, the user can interpret the response that the system gives in a different way than that the system (or the designer) meant.

With our 'Guess who you are talking to?' test, we evaluate the RM in an abstract manner. Participants interact with a response model in the terms of the response model, talking in terms of the answer frame. The user is his own Wizard of Oz, thus eliminating the first source of confusion. The RM responds in terms of the answer frame, without providing a natural language response. The system does not need to provide a response that might 'not really' fit the answer frame, for example because the conversation went in an unforeseen direction and the system has no appropriate response. We hope this eliminates the second source of confusion. This method comes at a cost. The abstract factors that the model uses and the personas that are portrayed by the model need to be explained to the participants. Also, there is a risk that participants may misunderstand the terms which creates a new possible source of confusion.

7.3 Experiment: Abstract Interaction

For our evaluation, 48 participants (42 male, mean age 24.8 with SD 3.7) volunteered to take part in the study. They were told they were going to interact with a system that could simulate a virtual suspect and that they were in the role of police officer interrogating this suspect. Their task was to determine with which of three personas

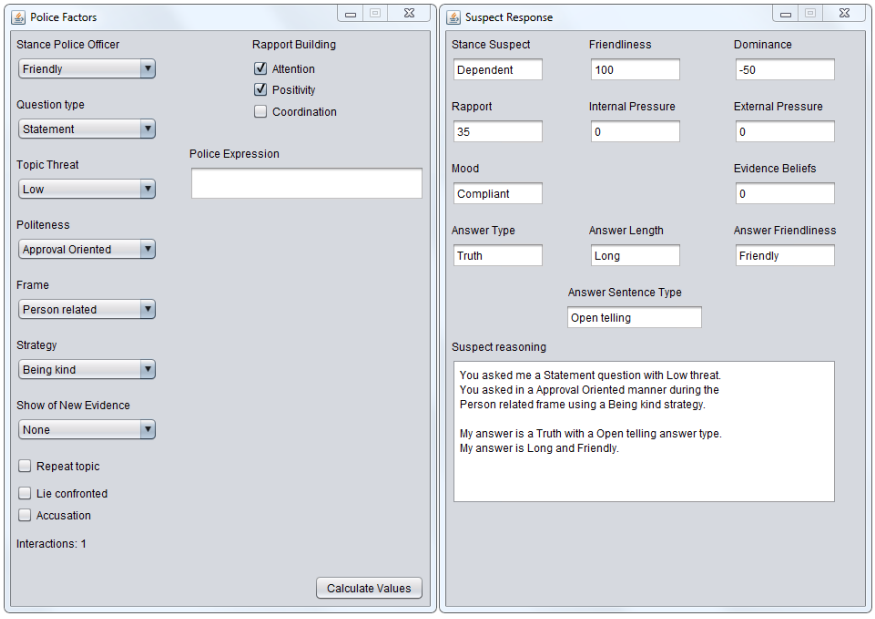

Figure 7.1: The Response Model Tester. The left panel shows the question frame that participants used to input their contribution to the conversation. The right panel shows the answer frame that the suspect used to convey his answer [208].

they were interacting. Each of these three personas were introduced in a small text, this text remained available for reference (see section 7.3.1).

Participants received an elaborate explanation of the variables in the RM. Each variable was explained and illustrated with several examples. The manner of interacting with the RM through the question frame and the answer frame was also thoroughly explained. Participants were encouraged to ask questions if something was unclear to them. Only when everything was understood could they start the experiment.

Each participant had two sessions of eight turns each with the response model: once with one of the personas and once with a random response generator (not based on a persona or response model). During each session they were asked to indicate with which of the personas they thought they were interacting. In addition, the participants were asked how confident they were about their choice and how realistic they found the interaction. After the interactions they were asked how familiar they were with the concepts and terms used in the response model and what their experiences were during the interaction.

Our Response Model Tester consisted of two graphical frames: the question frame where the participant could input his contribution to the interaction, and an answer frame which showed the response of the suspect (see Figure 7.1). All response model input and output, and the participant's choices, confidence, and realism ratings were logged.

7.3.1 Personas

The personality of the suspect in the RM can be set to reflect different personas. Three personas were created, based on personas from the DPIT-corpus [31, 141] prior to the experiment. These texts describing the personas were available to the participants:

> **Huls:** Mr. Huls is a friendly and mild family man. Recently he got into debt as he has no work. He takes this as a personal failure towards his family, he feels guilty for failing them. He is emotional and considers the feelings of others important.
>
> **Remerink:** Mr. Remerink married a wealthy man and holds his high social status in high regard. He is helpful when treated with respect, but gets very upset when disrespected. He perceived his arrest as an insult.
>
> **van Bron:** Mr. van Bron has a criminal record of drug-related crimes, assault, nuisance, and failure to comply with police requests. Has a history of abuse, neglect, and was raised in different foster care homes and boarding schools. He prefers to resolve situations with a big mouth and is prone to violence.

Participants interacted with one of these three personas or a random generator that provided random *answer frame* output. The personas were defined as follows in the RM as personality settings. These RM settings were *not* available to the participants.

> **RM setting Huls:** Dependent personality. High affiliation, sensitivity to rapport, and sensitivity to internal and external pressure. Low attitude to opposed.
>
> **RM setting Remerink:** Friendly personality. High dominance. Other variables moderate.
>
> **RM setting van Bron:** Aggressive personality. High dominance, attitude to being opposed, and sensitivity to internal pressure. Low affiliation, sensitivity to rapport, and sensitivity to external pressure.

7.3.2 Results and Discussion

In total 39 participants (81.25%) guessed correctly with which persona they were interacting. Participants who were correct were (statistically significant: $Z = -2.001, p < 0.1$) more confident (4.41) compared to the participants who were incorrect (3.67) (rated on a 5-point Likert scale (1=strongly disagree, 5=strongly agree)). The rating of realism was not significantly different: 3.90 for correct compared to 3.89 for incorrect. In the interactions where the responses of the system were random we might expect that each of the personas would be chosen an equal number of times (33%). However, the distribution of choices for the personas was Remerink 62.5%, van Bron 20.8%, and Huls 16.7%. Remerink was chosen significantly more often ($p < 0.05$) when participants interacted with the random generator. The average confidence level for interactions with personas was significantly higher 4.27 ($SD = 0.76$) compared to 3.46 ($SD = 0.77$) for interactions with the random generator ($Z = -4.2, p < 0.001$). The average level of realism for personas was significantly higher 3.90 ($SD = 0.52$) compared to 3.35 for random rounds ($SD = 0.89$) ($Z = -3.7, p < 0.001$).

After the experiment, we asked participants about their experiences during the experiment. People who interviewed the random generator *first* reported that they started doubting their first choice for a persona after they had interacted with the

second persona. They felt more confident about choice for the second persona and felt the first to be more random after they had interviewed the second. They reported that the second persona better met their expectations of one of the three personas. Some participants reported that when they had chosen a persona for the random output they felt they could not pick that persona again at their second run. They felt this way because the output was different from the first and they did feel some sort of confidence about their first choice. This led to some people choosing the wrong persona because they did not want to pick one they had chosen earlier. People tended to base their decision on parts of the output generated by the persona, they did not always look at all the output. They tried to rationalize 'weird random output' and actively tried to find reasons to fit it with their hypothesis and consider it realistic. We asked on which aspects of the suspect's responses they based their decision. Most participants based their choice only on parts of the suspect's responses. However, participants focussed on different parts and across all participants all of the answer frame output was used.

7.3.3 Conclusion Abstract Evaluation

The results of our 'Guess who you are talking to' test give an indication that our response model generates responses to user actions in such a way that the user is able to recognize a persona. This gives evidence of the validity of the response model and shows promise that the model can be used in the implementation of a believable virtual suspect character with various personality characteristics.

The method of evaluation of response models gives insight into the consistency with which a response model can portray a personality. It provides hints for improvements of the response model. Investigating which aspects of the model's response participants that 'guessed wrong' focussed on can provide hints on which aspects of the model should be improved. Another option to investigate how each part of the response model's response contributes to a 'correct guess' of participants is showing only some parts to different participants and comparing their 'correct guess' scores.

7.4 Natural Language Interaction Evaluation

Using the 'Guess who you were talking to' test, we found that the personality of the suspect was classified correctly 81.25% of the time by the participants, showing the 'error' of the RM alone being 18.75%. In this section we will investigate how the virtual suspect fared when participants had to use natural language in the interaction with the suspect. We expected the accuracy with which participants can 'Guess who they were talking to' would decrease as there were more potential sources for confusion.

7.4.1 Response Model

To reiterate, in chapter 5 [31] we analysed videos of police officers practising interrogations and defined several interpersonal, psychological, and linguistic concepts which are necessary to understand what goes on during an interview, including the concepts of *interpersonal stance* [116], *face* [28], and *rapport* [193] and the concepts

information and *strategy*. The RM discussed in chapter 6 is rule-based and the rules are based on these psychological theories and concepts. The implementation consists of four components: the *personality* of the suspect persona; a *Question Frame* that describes the question of the interviewer; the *Interpersonal State* as 'felt' by the suspect; and an *Answer Frame* that holds a description of the answer of the suspect (see Figure 7.2 top). The Question Frame influences the interpersonal state of the RM, taking into account the personality of the persona, and the 'current' Interpersonal State. The Answer Frame depends on this (updated) interpersonal state, the Question Frame, and the Personality, see Figure 7.2. For example, a persona with a friendly personality does not immediately become aggressive when confronted with an unfriendly question but if it is repeatedly confronted with unfriendly behaviour it can become aggressive.

7.4.2 Behaviour Realisation

We used components from the Virtual Human Toolkit [83] to build the virtual suspect. Specifically, we used the NPCEditor [118], a statistical text classifier that provides question-answer matching. It uses information retrieval techniques to match the user's input with a 'known' question and return the answers that are paired with this question. The questions and answers were authored by the authors and based on observations of many (practice and real) police interviews. All answers in the NPCEditor were annotated in terms of the *answer frame* of the RM. The NPCEditor provided several appropriate answers to a question of the user. A wizard interpreted the user's questions in the terms of the *question frame* of the RM. This triggered an update of the RM state. From the answers provided by the NPCEditor, the answer which annotation matched *interpersonal state* and *answer frame* state of the RM best was selected (see Figure 7.2). For example, if the RM was in a 'good mood' it selected a 'friendly' instead of an 'unfriendly' answer. The selected answer was sent to the VHToolkit Renderer that realised the behaviour. For all personas we used model 'Brad' from the VHToolkit, the voice of one of the authors, and the same NPCEditor script. The only difference between the personas was the setting of the personality in the RM.

7.4.3 Experiment: Virtual Suspect William

We asked 42 participants (age $M = 28.3$, $SD = 9.4$, 12 female) to interact with William, our virtual suspect, see Figure 7.3. There were four conditions, the RM personality of William was set to the personality of one of the three personas or the RM was a random answer frame generator, see for example interactions Table 7.1. The session started with an explanation on how to interact with the virtual suspect (see section 7.4.5). Participants interviewed the suspect until they completed their task: get him to say the name of an accomplice. Afterwards, they had to 'Guess who they were talking to' and filled out a questionnaire on how they had perceived the virtual suspect.

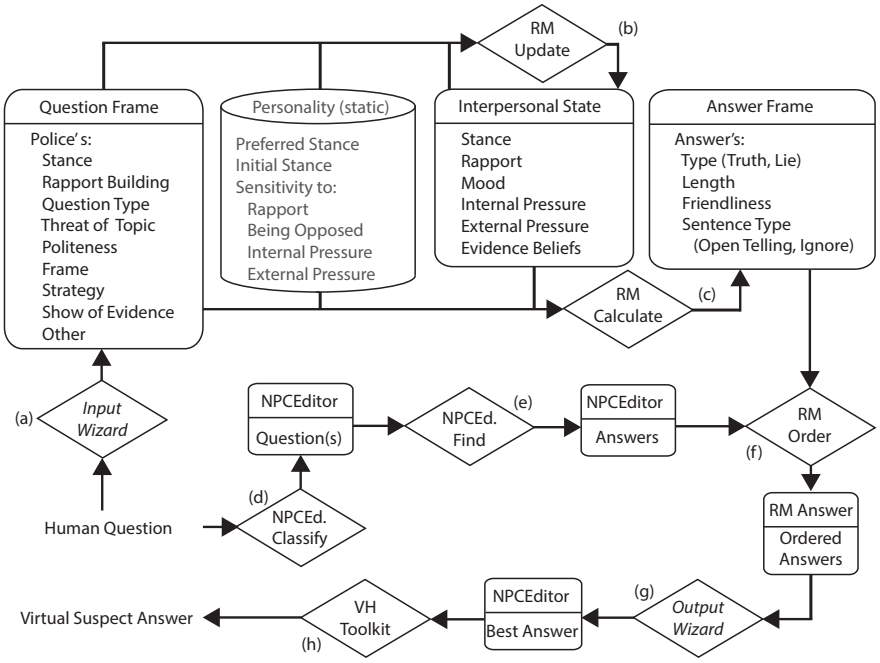

Figure 7.2: A description of the RM (section 7.4.1) and its integration in the VHToolkit (section 7.4.2). The user asks a question. A wizard provides values of the question frame (describing the question of the user) to the RM (a). The RM has a static personality (the persona representation). It has an interpersonal state (holds values describing how the suspect 'feels') that updates when a question frame is presented (b). The output of the RM is the answer frame (describing the answer the suspect will give) and it is calculated based on the question frame, personality, and (new) interpersonal state (c). Please refer to [31, 34, 35] for details on the terms in the RM. The NPCEditor finds appropriate answers to the human's question (d, e) and the RM orders those answers based on the answer frame (f). The answer that the RM selected to be most appropriate is executed by the VHToolkit (h). A wizard had the option to deviate from the RM suggestion and select a different answer if the NPCEditor selected answers that are inappropriate for the scenario (g), for example by misclassifying the question asked.

7.4.4 Case

The following case description, which resembles a police report, was provided to the participants:

> *William* is a suspect in a drug smuggling case. He was observed by a team of detectives delivering a suitcase filled with 20,000 XTC pills to the airport. He left the suitcase with suspected accomplice Shannon. Shannon was arrested with the drugs in her possession. This is proven and the suspect does not need to make statements about this. The house of the suspect was searched by detectives. In an office on the second floor a desk was found. This desk had a locked top drawer. A photo of Shannon was found in this drawer. It is not proven that this photo

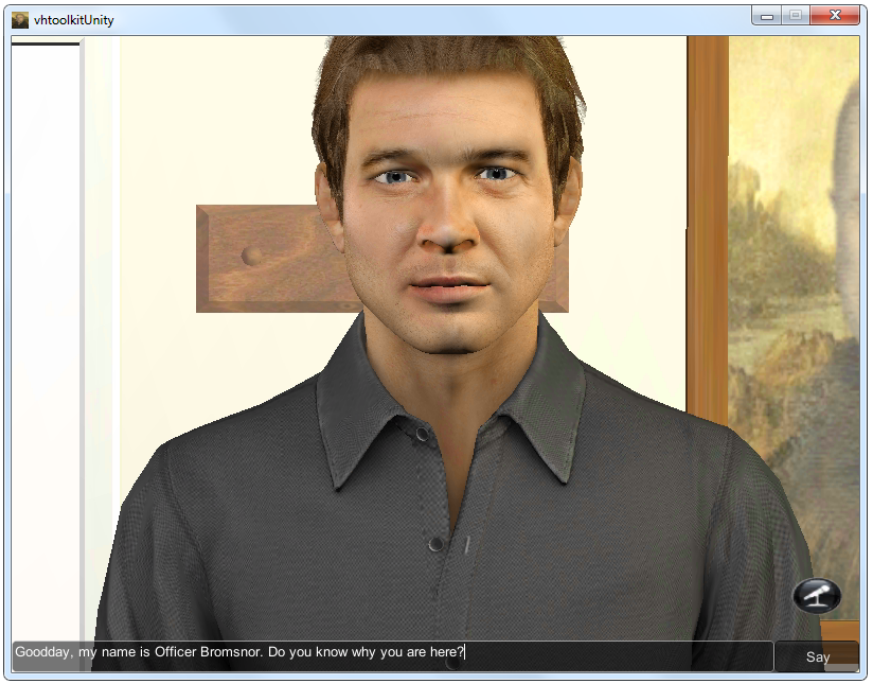

Figure 7.3: Virtual Suspect William portrayed by Brad from the VH Toolkit [83].

belongs to the suspect. It is not proven that Shannon and William know each other.

The police create an interview plan when they prepare for an interview. They determine the topics they want to address during the interview based on the tactical clues they have and they prepare questions for each of these topics. We prepared an interview plan and participants were told to follow it during their interview of the virtual suspect. The interview was over when William admitted to knowing Shannon, which was the inevitable eventual outcome of the interview.

7.4.5 Interacting with the Virtual Suspect

Participants had to follow the interview plan, but we explicitly encouraged them to add 'social padding' to the questions in the interview plan and make their contributions as natural as possible. The participants had to type their contribution to the conversation, when satisfied with the contribution press *ENTER*, and then pronounce their contribution in the way it was meant. The virtual suspect would respond based on *what* the participant typed and *how* they said it. The written contribution (the 'what') was processed automatically by the NPCEditor and the social spoken contribution (the 'how') was interpreted by a wizard (see Figure 7.2). The contribution participants typed had to be what we called a 'complete contribution'. This meant that it should include something for the suspect to respond to like a question or a

Table 7.1: Example Q&As for the three personas in different phases of interaction. The officer's question is in italic and the virtual suspect's answer is shown below it for each of the personas. Q1 and Q2a are asked at the beginning of the interaction, showing the effect of the initial RM status of the persona on the answers A1 and A2a. Q2b is asked after a pleasant conversation in which the police officer managed to build rapport etc. A2b shows the answers for each of the personas after this pleasant interaction. Q2c is asked after an unpleasant conversation where the officer was unfriendly and intimidating. A2c shows the effect this has on the answers of the suspect personas.

Q&A	**Huls**	**Remerink**	**van Bron**
Q1		*Where do you live?*	
A1	I'm living at Mainstreet 12 in Venice.	Why should I tell you where I live, didn't you guys just arrest me at my place? Go figure it out you dumbass!	I live on the moon, I'm actually from Mars.
Q2a		*Do others use your desk?*	
A2a	Well, sometimes when I have guests they sleep in the office. I guess they might use the desk when they are in there.	Access smackses! It's my desk. No one got any business there. No one gets access... get it? Smackses!	Access smackses! It's my desk. No one got any business there. No one gets access... get it? Smackses!
	Interaction where the officer is building rapport, being friendly, etc.		
	↓		
Q2b		*Do others use your desk?*	
A2b	Well, sometimes when I have guests they sleep in the office. I guess they might use the desk when they are in there.	Well, sometimes when I have guests they sleep in the office. I guess they might use the desk when they are in there.	I guess when I have guests they could use the desk.
	Interaction where the officer is intimidating, unfriendly, face threatening etc.		
	↓		
Q2c		*Do others use your desk?*	
A2c	What ever. It's like the public library in my office. The whole neighbourhood uses my desk.	Access smackses! It's my desk. No one got any business there. No one gets access... get it? Smackses!	Access smackses! It's my desk. No one got any business there. No one gets access... get it? Smackses!

statement. For example, *'OK.'* would not be a complete contribution but *'OK, but what else can you tell me about your office?'* would be. The virtual suspect responded when the participant finished pronouncing his or her sentence. Alternatively, it could occur that the virtual suspect was unable to understand the participant's sentence. In this case the suspect would interrupt after the participant pressed *ENTER* and said *'What do you mean?*'. This meant the participant had to change the written contribution and try again. We gave written and oral explanations and gave ample opportunity for questions. During the start of the interview we provided a reminder of the in-

teraction procedure if necessary. All participants understood the procedure and had a meaningful interaction with the virtual suspect. After the interaction, participants received a description of the personas and had to choose which of the three personas they thought was most similar to William, report the confidence in their choice, and fill out a questionnaire about how they perceived William's personality.

7.4.6 Results

In total there were 42 participants of which 53.1% or 17 guessed correctly whom they were talking with resulting in $\kappa = 0.295$. This was better than chance (33.3%), but worse than [34]'s result of 81.25% correct. There was no correct answer for the 10 participants that interacted with a random generator. Overall, van Bron was recognized best: 60% of the RM acts of van Bron were perceived as van Bron (*recall*) and 66.7% of the people who thought they were interacting with van Bron were correct (*precision*). Remerink had a recall of 54.5% and precision of 46.2%, and Huls had a recall of 45.5% and precision of 50%, see Table 7.2.

The confusion about personas tells us something about the possible reason for the mistakes and thus how serious these mistakes are. From the descriptions of the personas Huls, Remerink, and van Bron we could argue that they increase in offensiveness and decrease in friendliness. Following this rationale we argue that Remerink is more similar to Huls and van Bron than Huls is to van Bron. This is also reflected in the data. Huls is mistaken for Remerink 6 times but never for van Bron. Remerink is mistaken for Huls 2 times and 3 times for van Bron. Finally, van Bron is perceived as Huls 3 times and as Remerink 1 time. If we consider the differences between personas as a step (e.g. the difference between Huls and Remerink is one step, but Huls and van Bron is two steps) we see that 12 out of 15 misclassifications are one step from the intended persona and only three are 2 steps. This tells us that the confusion is not random. Rather, the system is able to answer in an extremely unfriendly manner (which is necessary to act as van Bron) but can do this even when it acts as Huls when the user is very unfriendly and gets Huls angry (or when the system has no friendly answers available).

The random setting for the RM provided random *Answer Frame* output. There is no correct answer for the 10 participants that interacted with the random generator.

Table 7.2: Table showing the relation between the RM personality setting (the persona it *acted*) and what persona the participants *perceived* most similar to the virtual suspect. It includes the totals for the RM settings and the totals for the perceived personas. For each persona it includes the accuracy of the perception (recall) and the accuracy of the RM (precision). Finally, the perceptions of the random interactions are presented.

	Acted (RM setting)					
Perceived	Huls	Remerink	van Bron	Total Perc.	Precision	*random*
Huls	**5**	2	3	10	50%	*0*
Remerink	6	**6**	1	13	46.2%	*8*
van Bron	0	3	**6**	9	66.7%	*2*
Total RM Setting	11	11	10	32		*10*
Recall	45.5%	54.5%	60%			

Table 7.3: The confidence the participants had in their choice for a persona.

	Huls	**Remerink**	**van Bron**	random
Mean	5,45	5,91	4,80	5,20
SD	0,82	1,22	0,92	0,79

In this condition, the content of the answer was appropriate but the interpersonal form was random. We might expect a uniform distribution of choices of personas. However, Remerink was chosen 8 times, van Bron 2 times, and Huls never, see table 7.2. Possibly people were confused by the inconsistency of the behaviour as the suspect could for example go from friendly to unfriendly and back every turn. Remerink might be the persona that fits such behaviour best. From the Remerink description: "He is *helpful* when treated with respect, but gets very *upset* when disrespected". This makes explicit that he is capable of a wide range of interpersonal behaviours, perhaps wider than the other two personas. The random responses are very likely to include at least some unfriendly or aggressive responses which might explain why Huls was never chosen. Also, the random responses are unlikely to be only unfriendly and aggressive which is what participants might have expected from van Bron. This might explain the lower number of choices for van Bron.

The confidence observers have in their 'Guess who you were talking to' choice tells us something about the clarity of the persona acts of the response model. If the virtual suspect displays confusing behaviour it is likely that participants are less certain about their choice. Participants answered on a 7-point scale how confident (lowest (1) or highest (7)) they felt about their choice. We expected the confidence to be lower when the responses of the virtual human lack clarity as they do in the random condition. Indeed, we found that the confidence in choice for each of the RM settings (the three personas and the random) differs close to significance level, (Kruskal-Wallis) $\chi^2 = 7.532, p = 0.057$. However, people who interacted with van Bron were less certain about their choice than people in other RM settings, where participants who interacted with Remerink were most confident in their choice, see table 7.3. Moreover, the difference in confidence was only significant (or approaching significance) for Remerink-random (Mann-Whitney $U = 30.0, p = 0.066$) and Remerink-van Bron ($U = 22.5, p = 0.018$), all other RM settings did not produce significant differences on confidence. So, our hypothesis that the random condition would result in lower confidence ratings holds true only when comparing random to persona Remerink. This is interesting because we earlier expected that the random condition was often interpreted as Remerink because Remerink was most likely to show a wide variety of behaviours. However, people that interacted with random were less certain about their choice than when they were interacting with Remerink. Note that this is regardless of whether participants were correct. When we look at the confidence of those that were correct the difference between RM persona-settings again differs almost significantly, $\chi^2 = 5.349, p = 0.069$. However, the confidence of participants that were incorrect does not differ significantly, $\chi^2 = 0.387, p > 0.5$. For the participants that were correct only the confidence between RM setting van Bron and Remerink differed significantly ($U = 7, p = 0.044$). It seems that van Bron

showed behaviour that made participants doubt their choice for him. This might be due to the volatile nature of his personality: he can be easily swayed from friendly to aggressive. Also, most participants were doing their best to be friendly and build rapport. This made even the nasty persona van Bron friendly if they persisted and participants might have been confused by their 'success' in turning him friendly towards the end of the interrogation.

7.4.7 Perception

Participants filled out a questionnaire after their interaction with the virtual suspect. The questionnaire consisted of 7-point scales in the form 'I thought William was ...', strongly disagree (1) and strongly agree (7). It consisted of 15 items asking how the virtual suspect was perceived. In addition there were two items investigating how well William understood the participant and how well the participant understood William. Of the means of answers to 17 items in the questionnaire the means of answers to 11 items differed significantly or close to significance ($p \leq 0.1$) for the different RM settings, see Table 7.4. These items show how the virtual suspect is perceived differently over the different RM persona settings.

Table 7.4: The means for the items on the perception scale that differed significantly over the RM settings. The significances for the difference between the RM settings (Kruskal-Wallis). (* for $p \leq 0.05$ and ** for $p \leq 0.01$).

	likeable	honest	competent	warm	credible	approach-able	sincere	friendly	confident	polite	innocent
random	4.600	3.900	3.500	3.400	4.000	5.200	3.500	3.400	4.600	3.200	1.400
v.Bron	3.500	5.100	3.500	2.800	3.700	3.300	3.700	2.400	4.700	2.200	2.200
Remer.	5.273	5.727	5.182	4.455	5.273	5.364	5.273	4.909	6.000	4.545	2.545
Huls	4.909	5.636	4.909	4.091	5.364	5.455	4.364	4.455	5.818	4.455	2.636
$\chi^2 =$	9.332	11.247	10.625	8.847	10.420	12.618	7.676	15.101	8.309	14.665	7.189
$p =$	0.025*	0.010**	0.014*	0.031*	0.015*	0.006**	0.053	0.002**	0.040*	0.002**	0.066

Table 7.5: The means for the items on the perception scale that differed significantly over the RM settings for the participants that were correct. The significances for the difference between the RM settings (Kruskal-Wallis) for the participants that correctly guessed with whom they were talking. (* for $p \leq 0.05$ and ** for $p \leq 0.01$).

	honest	competent	warm	informed	credible	approach-able	sincere	friendly	polite
van Bron	4.667	3.167	2.333	4.167	3.667	3.000	3.000	1.667	1.833
Remerink	5.833	5.667	4.167	6.000	5.333	4.833	5.167	4.667	4.167
Huls	6.200	4.800	4.400	5.000	5.200	5.600	5.400	5.000	4.600
$\chi^2 =$	4.808	8.816	6.394	5.833	5.392	7.409	6.714	11.564	9.409
$p =$	0.090	0.012*	0.041*	0.054	0.067	0.025*	0.035*	0.003**	0.009**

This means the virtual suspect was able to show different behaviour on these items. It seems likely that participants managed to differentiate between the personas because they could differentiate the behaviour of the virtual suspect on (at least) these items. This does not necessarily mean that participants who were able to distinguish between the personas interpreted the behaviour of the virtual suspect differently on

these items. In fact, the items that differed significantly on the RM settings for only those participants that got it correct are other items than all participants' ratings. Table 7.5 shows the significance levels for the items that differed significantly or close to significance ($p \leq 0.1$) for correct participants. Comparing Tables 7.4 and 7.5 shows that the items 'likeable', 'confident', and 'innocent' did not differ significantly for the participants that were correct. Additionally, item 'informed' differed significantly only for the participants that were correct.

The items that asked about the ease of communication and clarity did not differ. In fact, items 'William understood what I said' ($M = 4.667, SD = 1.356$) and 'I understood what William said' ($M = 6.238, SD = 0.878$) seem to indicate that participants had no problems communicating with the agent.

7.5 Discussion

Getting the behaviour of a virtual character right is not easy. Getting the persona right is an important step towards a believable virtual suspect that can be used to train police officers to interrogate suspects. The RM calculates an answer frame; interpersonal features of responses of a virtual suspect, based on a persona and the question asked. In the first experiment we attempted to isolate the performance of the RM; how well its responses could be interpreted as belonging to one of three personas. There participants interacted using question frame values as input and received answer frame values as output. In the second experiment we expanded on those results and investigated what effect using a virtual agent that can understand and use natural language in the interaction has. In the abstract interaction, participants were able to guess correctly with which persona-setting in the RM they were interacting in about 80% of the time. This would indicate that the RM leads to confusion about who the RM is trying to enact in about 20% of the participants. In the interaction with natural language we found the accuracy of the 'Guess who you were talking to' test decreased to about 53%, showing the influence of natural language in the interaction and the importance of good authoring of responses for a virtual human. Other possible reasons for the decrease in performance include the appearance and voice of the virtual suspect.

We found that the personas that differed most were less likely to be confused. This means the RM was indeed able to select different behaviour for different personas and that the behaviour differed more when the personas were more different. So, it appears that confusion was not random. In fact, we argue that some participants managed to change a persona's initial mood and overcome its personality so that it showed behaviour not characteristic for the persona. The ability of the RM to do this is what caused the confusion. In police trainings this is exactly what the (virtual) suspect actor must do: respond to the behaviour of the trainee. The virtual suspect William is not yet able to provide a reflection after the interaction that tells the participant how (un)successful they were at changing the 'mood' of the suspect. We expect participants would have been more accurate at the 'Guess who you were talking to' test if they had had such information.

To create a virtual suspect that requires no wizard and that is capable of having a more natural interaction, we need to include automatic recognition of speech and

the interpersonal features of speech. Also, the system will have to be able to give feedback on the interaction in terms of the RM to facilitate reflective learning. These issues remain for future work.

Part IV

Conclusion

8

Discussion and Future Work

In this chapter I will discuss the contribution of our work on building an interactive virtual suspect that can be used in social skills training for police interrogations. I will reflect on the challenges that remain and the opportunities that tackling these challenges will provide.

In the first chapter I asked myself why suspects in a police interrogation say what they say and how they say it. To investigate this, we collected an audio-visual corpus of practice police interrogations. We investigated which psychosocial theories observers use to describe the social dynamics of the interaction. Our analyses showed that theories about interpersonal stance, face and politeness, and rapport are necessary to describe the social aspects of an interrogation. Additionally, concepts of information exchange and strategy are important. My goal was to use this insight to build a model that can make a virtual suspect respond like a human suspect. Such a virtual suspect can be interrogated in a serious game by a student of the police academy to learn how to conduct a proper police interrogation.

8.1 Annotating Behaviour

We saw in chapters 3 and 4 that raters asked to annotate interpersonal stance often did not agree. Labelling interpersonal stance is hard and something fundamentally different from, say, counting the number of tags that children make in a game of tag [131]. The amount of interpretation that the annotator has to do is crucial in understanding this difficulty and the amount of interpretation that is necessary depends on the type of content. Potter and Levine-Donnerstein [150] distinguish three types of content that can be annotated:

1 Manifest content can be observed directly without interpretation, it is that which is on the surface: for example, counting the number of tags in a game of tag.

2 Pattern latent content has to be inferred from surface features, but cannot be ob-

served directly. An example is an (unobservable) illness that can be diagnosed from visible symptoms. Here a good annotation schema can provide rules determining which combinations of manifest content should lead to an annotation of a pattern latent class.

3 Projective latent content lacks such clear relations to easily quantifiable features. Annotations of this category rely on, what Potter and Levine-Donnerstein call, the annotator's mental schemas. It relies on concepts that most people share and understand on an intuitive level, but that are extremely difficult or impossible to exhaustively define. One example of a projective annotation task is rating the dominance or friendliness of a person.

We believe it proved difficult for observers to agree on the annotation of interpersonal stance because interpersonal stance refers to projective latent content and because utterances can be taken in different ways. Exemplary for this was a remark one of the raters in chapter 3 made during a discussion about a particular utterance by the officer that was annotated differently by about half of the group. About half of the raters thought the behaviour was friendly and annotated it as such, while the other half rated differently, they considered the utterance to be aggressive behaviour. One of the raters with a different annotation said, "*I saw the police officer was annoyed by the suspect and got angry, but he did not show it because he was taught to remain friendly, so I annotated he had an aggressive stance because this was his actual stance*". He behaved in a friendly manner yet he was angry. Her annotation was based on a 'gut-feeling' or intuition: the annotated stance was *not* displayed by the police officer but inferred from the context. All raters could understand both annotations after the discussion. The confusion comes from what was to be annotated, the projective or the pattern latent content point of view.

We found that the disagreement between annotations of interpersonal stance was not random, but that when annotators disagree in their choice they often choose labels that are next to each other in Leary's Rose. Based on the reliability analyses in chapter 3 we conclude that interpersonal stance is a viable model to make sense of the way people take a stance towards each other in a social encounter. The question as to what causes this confusion between raters remains. Reidsma [156, p19-20], in his PhD thesis, suggests six sources of rater disagreement:

1) "Inadequate selection of relevant concepts for inclusion in the annotation scheme;

2) Invalid or imprecise annotation schemas;

3) Insufficient training of the annotators;

4) Clerical errors;

5) Genuinely ambiguous expressions;

6) A low level of inter-subjectivity."

We suggested that interpersonal stance is a fuzzy concept and as such it might be impossible to obtain high interrater agreement. The fuzzy nature of interpersonal

stance means that some behaviour can be interpreted in more than one way. In some cases interpersonal stance might be genuinely ambiguous (item 5 of the causes for confusion that Reidsma [156] proposed). Poesio and Artstein [148] argued that disagreements in annotation, of ambiguous anaphoric relations in linguistics, does not mean that the annotation scheme is faulty. However, to prevent such confusion, we might argue that the annotation schema could provide a clearer instruction on what to annotate. One option would be to instruct raters to annotate the behaviour that was displayed overtly, another would be to interpret the stance that the police officer or suspect 'feels'. This might address the issue when confusion was due to the impreciseness of the annotation schema (item 2 of Reidsma's causes for confusion).

We are interested in the internal state of the suspect, because we wish to use the information from the annotations to create a computational model of the internal state of a virtual suspect. This means that we have to face the disconnection between the interpretation of the shown behaviour and the 'feelings' that motivated this behaviour. It may not always be apparent to all raters when there is a possibility to interpret the behaviour in this projective manner. Some raters might pick up on some subtle cue and interpret behaviour in a projective manner, others might inadvertently think there is some cue and annotate a latent feeling that was never there, while others might never see a reason to interpret the behaviour on any level and annotate only the manifest content of the behaviour. Thus, confusion might remain regardless of the instruction. It would appear there are two levels of inter-subjectivity in this case, the subjectivity of whether or not there is a projective latent interpretation to be made, and the subjectivity of what the interpretation in that case should be. Interestingly, interpersonal stance can be annotated with reasonable agreement when we take into account that raters often choose labels that are next to each other in the interpersonal circumplex. Interpersonal stance appears to be a fuzzy concept.

Training raters might have another potential issue. One option is to train raters extensively by having them discuss cases with low agreement, hoping they agree on annotations in new ambiguous cases that are somehow similar. However, another group of raters might arrive at a different (implicit) coding schema. Perhaps inter-rater reliability looses some of its use when extensive training is required to achieve high agreement. In particular, when the ratings will be used in an application context where some end-user needs to agree with the distinctions that were made. In other words, training raters might be bad for something that we suggest might just as well be called 'universal naive agreement': the agreement that untrained raters, members of the general public that to some extent do understand the annotation task, would have (see e.g. [91]).

We found that obtaining rater agreement on an utterance level was difficult. Yet, a majority vote in chapter 3 did reveal a pattern showing 'what is going on' in a police interrogation in terms of interpersonal stance taking. Crucially, this pattern appears only over longer interaction periods: where at first the suspect is predominantly withdrawn, then changes to a more cooperative stance, she finally turns to a competitive stance. Chindamo et al. [44] suggest "Communicative Stance = Attitude which, for some time, is expressed and sustained interactively in communication" [44, p618] and they go on, "The qualification 'for some time' means that normally a stance is

not short term but sustained through a sequence of contributions". Our findings corroborate their suggestion and further define 'some time': we found that the suspect's stance changed roughly every 150 turns. It should be noted that the duration between these stance changes in the interview were likely due to contextual reasons that can explain them. Our hypothesis is that inter-annotator agreement will rise when rating segments of around this length on interpersonal stance, something further research should investigate.

Related to this issue of universal agreement is the notion of 'golden truth'. In some cases it is possible to obtain a golden truth for what Potter and Levine-Donnerstein [150] call projective latent content. If we take 'the stance that a person experiences or intends to convey' as projective latent content, and we ask an actor to act out a stance, we have access to the stance as it was intended and felt. In chapter 4 we found a correlation between the stance as it was intended by the actor and the stance rated by independent observers. From this we conclude that observers are able to correctly annotate interpersonal stance.

The proficiency of the actor matters for rater-rater agreement and for the rater-actor agreement. Some actors are better than others in displaying behaviour that observers can *agree on*. Some actors are better than others at displaying behaviour in a manner that observers can *recognize* as the intended behaviour. Creating the behaviour of a virtual suspect based on an actor's performance should warrant an investigation into an actor's proficiency on these two matters. The behaviour of amateur actors was often labelled as 'artificial' in chapter 4, skewing the annotations of most fragments towards a submissive-hostile stance independent of the stance that was intended by the actor. Artificial is an adjective for a submissive-hostile stance. People show a mixture of behaviours that are related to stance taking and it depends on the perspective of the judge which aspects determine how the stance is perceived.

It is an open question what this means for the way people perceive the fabricated stance behaviour of an artificial human. Obviously, an artificial human can only show fabricated behaviour as it is a fabricated entity. Conversely, a human is expected to show 'real' behaviour that is not fabricated. So, when a human shows behaviour that can be interpreted as fabricated, people perceive it as artificial. Our hypothesis is that people will not rate the behaviour of a virtual human as artificial as they are aware that an artificial human can only show fabricated behaviour.

8.2 Response Modelling

Credible virtual human responses are crucial for a serious gaming system that will train police officers to do a proper interrogation. If players are to engage in a meaningful social interaction with an artificial agent, they need to suspend their disbelief in the realism of the artificial interaction first (e.g. [117, 138]). In gaming this suspense is aided by creating presence with a compelling narrative, realistic virtual environments, events, non-player characters, and the consistency between these. Perlin [146, p146] put it like this, "characters should be free to respond in a way that accords with their nature". We focussed on creating a system in which actions of the user will be met with credible responses of the virtual human: a response model. As

Ochs et al. put it “Credibility, [...], relies greatly on the notion of consistency ” [138, p281]. They go on to distinguish two dimensions of consistency in artificial characters (in games):

1 ‘Consistency with past behaviour’ which entails that the behaviour of an artificial character should match its personality and the events that preceded a ‘current’ behaviour.

2 ‘Consistency with the current environment’.

Our efforts towards this consistency took the form of a response model that can be set to a personality and takes into account the interpersonal events that have occurred in the interaction in an interpersonal state that is updated with new events. This state is used to compute an answer frame that abstractly describes the response of the artificial agent, see chapter 6. The results from our evaluation of this RM, in chapter 7, revealed that participants could recognize with which personality they were interacting: a sign of consistency.

Credible virtual agent behaviour leads to higher presence, and less need to suspend disbelief. Traditionally in virtual reality research the focus was on physical presence, which relates to the extent people feel that they are present in the virtual world (e.g. [215]). Social presence is the notion that a virtual being exists and can interact with you [84], even if that being is only human-like and only seems intelligent [21]. Lee and Nass [117] revealed that “users feel a stronger sense of social presence when the personality of synthesized voice matches the personality of textual content than when those two are mismatched”. Others have found that presence increases when the interactant is real or known [152], or when the avatar is self-designed [9]. This shows that credibility of virtual environment has many facets that can be optimised to increase a user’s presence. However, whether the increase of presence means an increase in learning is doubtful [214].

In a police interrogation, the police officer and the suspect have distinct roles. These roles are defined by external factors such as the power the police officer has in dictating when the interview starts and stops, yet there are also internal factors such as the personality of the interlocutors [58]. The external influences on roles are implicitly defined in our RM, for example the virtual agent is a suspect. For generalisability to other domains, it is necessary to explicitly include these external influences on role and relationship in the RM. The status that comes with a role is one of these influences that could be considered for explicit modelling. This can be done in a similar manner to how we gathered which factors are relevant to include in an RM: using observation of interaction in the domain. The challenge is to explicitly include external domain-specific factors in a model in such a way that they can “encode” for different domains.

8.3 Ethical Considerations

A system that can autonomously teach students social skills will change learning, undoubtedly there are many exciting opportunities with such technology. However, it is very important to investigate possible concerns about this technology: professional

training actors worry that they will be replaced, teachers worry they will lose influence over what and how they teach, and students worry they will not get the personal attention they need and deserve. Whether these concerns are valid will depend on how the technology will be embedded in the learning environment. We envision a future with the technology that looks like this: students of the police academy familiarise themselves with the theory of interpersonal stance (or Leary's Rose as they call it). With their fresh knowledge they can engage in interactions with a virtual suspect. Here they experience (or discover) the effects of their interpersonal stance taking on the responses of the virtual suspect. These experiences will illustrate the theory they were taught, so they get a higher level of interpersonal social skills before they interact with a professional training actor. Additionally, the interaction with the virtual suspect will give them insight into the topics they find difficult. In a classroom setting with the help of their teacher and with a professional training actor, the students will get the opportunity to put the finishing touches on their social skills. After their training is complete, police officers get the opportunity to brush up and maintain their skills by interacting with the virtual suspect whenever they want. One thing is apparent (to me), actors will not and *should not* be replaced, teachers will still be crucial in determining what and how they teach, and students will get an additional tool with which they can find out for which topics they should seek more training.

Another ethical issue that needs to be considered is the risk of stigmatisation. In any game scenario the virtual suspect will portray someone. This means that a virtual suspect has an appearance and that, like all humans, it will have an ethnicity and social status. Training officers with the same avatar might create a risk that it consolidates some prejudice that exists with the person, group, or organisation the avatar belongs to. Too few avatars, scenarios, voice actors, or types of dialogue in a training might create an (unconscious) bias towards some person or group in the user when this avatar is always 'playing the bad guy'. Obviously it is imperative that the police[1] is unprejudiced. Fortunately, a solution might be in the risk: students could be brought into situations where they (or their teachers) feel they might have a prejudice in order to experience their bias and be corrected if necessary. A system capable of such training has similarities to systems that are used for training users how to behave and interact with people from a different culture (e.g. [54]).

Artificial humans push us to ask important questions. For example, what does it mean to be human and what does it mean to be real? One interesting approach towards answering this question is that of Japanese roboticist H. Ishiguro. He created a robotic doppelgänger, a 'geminoid'. This device looks like Ishiguro and can mirror his movements. It is so lifelike it can invoke the feeling of being in the presence of a human being [2]. Another question is, does an artificial social entity that 'feels real' have rights? In humans, privacy of one's thoughts is a fundamental right. For a serious game we look in the 'thoughts' of the virtual human to offer the user an explanation for the behaviour of the virtual human. Perhaps this last question can be answered simply by saying that artificial humans have artificial human rights.

[1] This goes for any person, group, or organisation.

[2] See for example, http://spectrum.ieee.org/robotics/humanoids/hiroshi-ishiguro-the-man-who-made-a-copy-of-himself

8.4 Future Work

Validation of RM responses with the target user group and investigating the learning effects of an interaction with a virtual suspect are two of many possible future endeavours that are important for the adoption of virtual agent technology in social skills training. During informal testing with some instructors from the police academy and other law enforcement agencies, they showed willingness to suspend their disbelief and engage in a social interaction with the virtual suspect. However, they were still unconvinced that the virtual suspect could simulate a sufficiently rich interaction so that a student could develop feelings of rapport with the virtual suspect. Also, they were concerned whether a computer system would be able to observe and interpret the subtle social cues present in a police interrogation. These concerns should be addressed with a proper evaluation and validation of the technology, provided the technology is developed and available.

Efforts should be made towards creating a most natural and human-like virtual agent. For example, text-to-speech (TTS) still needs to improve. At this moment the realism of TTS is at a level that voice actors remain necessary to create a compelling illusion of presence. One way in which the perceived quality of TTS might be improved with current technology is creating a set of TTS-voices with the same voice actor[3], where the difference between the voices is the emotional state of the voice actor. The voice with the emotional state that is most appropriate, as indicated by an RM, should be selected to pronounce the utterance, thus increasing the realism of the virtual character.

8.4.1 Knowledge Representations and Reasoning

In chapter 1, we discussed that in police interrogations there is always a danger of obtaining a false confession. There are three general types of false confessions:

1 Voluntary false confessions;

2 Forced false confessions that the suspect does *not* believe;

3 Forced false confessions that the suspect believes.

One of the ways to ascertain whether a confession was false is comparing the information that a suspect provides with the confession to the knowledge the police had at the time of the interrogation. The assumption is that an interrogator might have (unwillingly) disclosed information about the case to a suspect and that the suspect (unknowingly) incorporated this information in his or her memory [202]. Virtual humans might play a role in training police officers to not obtain false confessions. For this the system should have the ability to reason about the knowledge that the interlocutors hold. Inspiration for systems with this capability can be found in research about theory of mind (TOM) (e.g. [210]).

[3]See for example the company Fluency, that offer to create a TTS voice with the characteristics of a speaker. Their service is intended for patients who, due to some disease, will lose their voice. http://www.fluency.nl/

8.4.2 Feedback

Reflective learning, as opposed to experience learning, requires feedback. Currently, the implementation of our RM can show the user a log of the variables in the RM over time during the interaction. This shows the changes in the virtual suspect's interpersonal state. Combined with the actions that the user made to cause each change in the RM this can be a powerful learning instrument. However, as Pereira et al. [145] noted, feedback about formative assessment is a challenge that must be tackled to improve serious games for training interpersonal skills. Linssen et al. [121] show that using such feedback to effectuate directly measurable learning gains is a feat that is not easily obtained, but interestingly suggest that a more difficult learning task might effectuate a learning effect.

8.4.3 Real-time Behaviours

In chapter 3 we saw that in a police interview turn-taking is not as neatly structured as Sacks, Schegloff, and Jefferson [165] suggested. We found that occurrences of overlap and silences are common and that turn-taking can carry meaning. Our prototype currently offers only turn by turn interactions, see chapter 7. Ideally, for a serious game the turn-taking between the user and the virtual suspect will be human-like, meaning that the interlocutors can interrupt each other and talk at the same time. Fighting for the turn can be a fight for the status and dominance. Also, being silent for a long time after a question can be a powerful tool to put pressure on a suspect. Being able to practise these techniques in a serious game would be a great addition to the skills that can already be trained with this technology. Ravenet et al. [153] recently reported work on an affective real-time turn-taking system. They created a system where multiple virtual agents can converse and display non-verbal behaviour, including turn-taking. This behaviour varies with the interpersonal attitude held by the agent. This work is a continuation of the work by Thórisson et al. [192]. Their Ymir Turntaking Model is a broad computational model of conversational skills but does not consider the expression of attitudes. However, for a serious game it is important that the affective behaviours are also tailored to the learning goals of the user [30]. To the best of our knowledge there is not yet a system that is capable of continuous verbal and non-verbal affective virtual human behaviour that consists of real-time response to a user, and that can adapt the virtual agent's attitudes to deliver an optimal training experience for the user.

8.4.4 Standardised Evaluation Paradigm for Social Systems

The final consideration in this thesis is inspired by a discussion at the Virtual Agents for Social Skills Training (VASST) workshop[4] at the INTETAIN 2016 conference. Here the point was raised that a standardized paradigm of evaluation is lacking for systems with social skills. A standard evaluation that would allow for objective comparison between socially aware systems regardless of the exact domain in which they are applied. Taking a look at how other fields evaluate and compare their work shows

[4] http://www.intetain.org/2016/show/vasst-2016

that evaluation paradigms for socially intelligent systems have a long way to go. See for example the NIST Speaker Recognition Evaluation (e.g. [76]) that allows for comparison of state-of-the-art speaker recognition systems. Challenges are updated to reflect the latest advances in the field: "The overarching objective of the evaluations has always been to drive the technology forward, to measure the state of the art, and to find the most promising algorithmic approaches"[5]. For serious gaming, an important aspect that a game should get right is *engagement* [26]. Perhaps measuring the engagement that users experience while interacting with a social system could be a starting point for a standardized evaluation paradigm.

[5]From: http://www.itl.nist.gov/iad/mig/tests/spk/, accessed July 2016.

Bibliography

[1] ALLAN, K. *Natural language semantics*. Blackwell Publishers Ltd., Oxford, 2001.

[2] ALLWOOD, J., CHINDAMO, M., AND AHLSEN, E. On identifying conflict related stances in political debates. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)* (2012), pp. 918–925.

[3] ANDERSON, K., ANDRÉ, E., BAUR, T., BERNARDINI, S., CHOLLET, M., CHRYSSAFIDOU, E., DAMIAN, I., ENNIS, C., EGGES, A., GEBHARD, P., JONES, H., OCHS, M., PELECHAUD, C., PORAYSKA-POMSTA, K., RIZZO, P., AND SABOURET, N. The tardis framework: Intelligent virtual agents for social coaching in job interviews. In *Advances in Computer Entertainment* (2013), Springer, pp. 476–491.

[4] ARGYLE, M. *Bodily Communication*, 2nd ed. ed. Methuen, 1988.

[5] ARTSTEIN, R., AND POESIO, M. Inter-coder agreement for computational linguistics. *Comput. Linguist. 34*, 4 (Dec. 2008), 555–596.

[6] AUSTIN, J. L. *How to do things with words*, vol. 1955. Oxford University Press, Oxford, 1975.

[7] AYLETT, R., VALA, M., SEQUEIRA, P., AND PAIVA, A. FearNot! – An Emergent Narrative Approach to Virtual Dramas for Anti-bullying Education. In *International Conference on Virtual Storytelling* (2007), vol. LNCS 4871, Springer Berlin Heidelberg, pp. 202–205.

[8] AYLETT, R. S., LOUCHART, S., DIAS, J., PAIVA, A., AND VALA, M. FearNot! – An experiment in emergent narrative. In *Intelligent Virtual Agents* (2005), vol. 3661 LNAI, Springer Berlin Heidelberg, pp. 305–316.

[9] BAILEY, R., WISE, K., AND BOLLS, P. How avatar customizability affects children's arousal and subjective presence during junk food-sponsored online video games. *CyberPsychology & Behavior* (2009).

[10] BALES, R. F. *Personality and interpersonal behavior.* Holt, Rinehart & Winston, 1970.

[11] BALLIN, D., GILLIES, M., AND CRABTREE, B. A framework for interpersonal attitude and non-verbal communication in improvisational visual media production. In *1st European Conference on Visual Media Production (CVMP)* (2004), IEE, London, UK.

[12] BÄNZIGER, T., MORTILLARO, M., AND SCHERER, K. Introducing the geneva multimodal expression corpus for experimental research on emotion perception. *Emotion 12(5)* (2012), 1161–1179.

[13] BÄNZIGER, T., AND SCHERER, K. R. Using actor portrayals to systematically study multimodal emotion expression: The gemep corpus. In *Affective computing and intelligent*

interaction. Springer, 2007, pp. 476–487.

[14] BARANYI, P., AND CSAPÓ, A. Definition and synergies of cognitive infocommunications. *Acta Polytechnica Hungarica 9*, 1 (2012), 67–83.

[15] BATESON, G. A theory of play and fantasy. *Psychiatric Research Reports 2*, 39 (1955), 39–51.

[16] BENNEWORTH, K. Police interviews with suspected paedophiles: a discourse analysis. *Discourse & Society 20*, 5 (2009), 555–569.

[17] BEUNE, K., GIEBELS, E., AND SANDERS, K. Are you talking to me? Influencing behaviour and culture in police interviews. *Psychology, Crime & Law 15*, 7 (2009), 597–617.

[18] BEUNE, K., GIEBELS, E., AND TAYLOR, P. J. Patterns of interaction in police interviews: The role of cultural dependency. *Criminal Justice and Behavior 37*, 8 (August 2010), 904–925.

[19] BEVACQUA, E., MANCINI, M., AND PELACHAUD, C. A listening agent exhibiting variable behaviour. In *Intelligent Virtual Agents* (2008).

[20] BICKMORE, T. Framing and interpersonal stance in relational agents. In *Functional Markup Language Workshop at AAMAS 2008* (2008).

[21] BIOCCA, F. Chapter 6 The Cyborg's dilemma. Progressive embodiment in virtual environments. *Journal of Computer-Mediated Communication 3*, 2 (1997).

[22] BIRDWHISTELL, R. *Kinesics and Context*. University of Pennsylvania Press, 1970.

[23] BLAAUW, J. A. *De Puttense moordzaak: de volledige geschiedenis van Nederlands grootste gerechtelijke dwaling*. De Fontein, 2009.

[24] BLASCOVICH, J., LOOMIS, J., BEALL, A. C., SWINTH, K. R., HOYT, C. L., AND BAILENSON, J. N. Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry 13*, 2 (2002), 103–124.

[25] BLUME, A., AND BOARD, O. Intentional vagueness. *Erkenntnis 79*, 4 (2014), 855–899.

[26] BOSSE, T., AND GERRITSEN, C. Towards Serious Gaming for Communication Training-A Pilot Study with Police Academy Students. In *Conference on Intelligent Technologies for Interactive Entertainment (INTETAIN)* (2016).

[27] BOUD, D., KEOGH, R., AND WALKER, D. *Reflection: Turning experience into learning*. Nichols Publishing Company, New York, 1985.

[28] BROWN, P., AND LEVINSON, S. C. *Politeness: Some universals in language usage*. Cambridge University Press, Cambridge, 1987.

[29] BRUIJNES, M. Affective conversational models: interpersonal stance in a police interview context. In *Humaine Association Conference on Affective Computing and Intelligent Interaction* (2013), pp. 624–629.

[30] BRUIJNES, M., KOLKMEIER, J., OP DEN AKKER, H., LINSSEN, J., THEUNE, M., AND HEYLEN, D. Keeping up stories: design considerations for a police interview training game. In *Proceedings of the Social Believability in Games Workshop (SBG2013)* (Enschede, The Netherlands, 2013), CTIT, University of Twente, p. 14.

[31] BRUIJNES, M., LINSSEN, J., OP DEN AKKER, R., THEUNE, M., WAPPEROM, S., BROEKEMA, C., AND HEYLEN, D. Social behaviour in police interviews: Relating data

to theories. In *Conflict and Multimodal Communication: Social Research and Machine Intelligence* (Switzerland, 2014), Springer International Publishing.

[32] BRUIJNES, M., OP DEN AKKER, R., HARTHOLT, A., AND HEYLEN, D. Virtual suspect william. In *Intelligent Virtual Agents* (2015), Springer, pp. 67–76.

[33] BRUIJNES, M., OP DEN AKKER, R., SPITTERS, S., SANDERS, M., AND FU, Q. The recognition of acted interpersonal stance in police interrogations and the influence of actor proficiency. *Journal on Multimodal User Interfaces* (2015), 1–24.

[34] BRUIJNES, M., WAPPEROM, S., OP DEN AKKER, H., AND HEYLEN, D. A method to evaluate response models. In *Proceedings of the 14th International Conference on Intelligent Virtual Agents* (Switzerland, 2014), T. Bickmore, S. Marcella, and C. Sidner, Eds., vol. 8637 of *Lecture Notes in Computer Science*, Springer International Publishing, pp. 67–70.

[35] BRUIJNES, M., WAPPEROM, S., OP DEN AKKER, H., AND HEYLEN, D. A virtual suspect agent's response model. In *Fourteenth International Conference on Intelligent Virtual Agents (IVA 2014); Proceedings of the Workshop on Affective Agents* (2014), L. Ring, Y. Leite, and J. Dias, Eds., Gaips Intelligent Agents and Synthetic Characters Group, pp. 17–24.

[36] BURKETT, C., KESHTKAR, F., GRAESSER, A. C., AND LI, H. Constructing a personality-annotated corpus for educational game based on leary's rose framework. In *FLAIRS Conference* (2012).

[37] BUSSO, C., AND NARAYANAN, S. Recording audio-visual emotional databases from actors: A closer look. In *Second International Workshop on Emotion: Corpora for Research on Emotion and Affect, International conference on Language Resources and Evaluation (LREC 2008)* (2008), pp. 17–22.

[38] CAFARO, A., VILHJÁLMSSON, H. H., BICKMORE, T., HEYLEN, D., JÓHANNSDÓTTIR, K. R., AND VALGARDHSSON, G. S. First impressions: users' judgments of virtual agents' personality and interpersonal attitude in first encounters. In *Intelligent Virtual Agents* (2012), Springer, pp. 67–80.

[39] CALLEJAS, Z., RAVENET, B., OCHS, M., AND PELACHAUD, C. A computational model of social attitudes for a virtual recruiter. In *13th International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2014* (2014), vol. 1, International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), pp. 93–100.

[40] CAMPOS, H., CAMPOS, J., MARTINHO, C., AND PAIVA, A. Virtual agents in conflict. In *Intelligent Virtual Agents* (2012), pp. 105–111.

[41] CARNEY, D. R., HALL, J. A., AND LEBEAU, L. S. Beliefs about the nonverbal expression of social power. *Journal of Nonverbal Behavior 29*, 2 (2005), 105–123.

[42] CASSELL, J., GILL, A. J., AND TEPPER, P. A. Coordination in conversation and rapport. In *Proceedings of the workshop on Embodied Language Processing* (2007), Association for Computational Linguistics, pp. 41–50.

[43] CHEESEMAN, P. Probabilistic versus fuzzy reasoning. In *Uncertainty in Artificial Intelligence Annual Conference on Uncertainty in Artificial Intelligence (UAI-85)* (Amsterdam, NL, 1985), Elsevier Science, pp. 85–102.

[44] CHINDAMO, M., ALLWOOD, J., AND AHLSEN, E. Some suggestions for the study of stance in communication. In *Privacy, Security, Risk and Trust (PASSAT), 2012 Inter-*

national Conference on and 2012 International Confernece on Social Computing (SocialCom) (2012), pp. 617–622.

[45] CHOLLET, M., OCHS, M., AND PELACHAUD, C. Interpersonal stance recognition using non-verbal signals on several time windows. In *Proceedings Workshop Affect, Compagnon Artificiel, Interaction* (November 2012).

[46] CHOLLET, M., OCHS, M., AND PELACHAUD, C. Mining a multimodal corpus for non-verbal behavior sequences conveying attitudes. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (may 2014), European Language Resources Association (ELRA).

[47] CLÉMENT, S., VAN DE PLAS, M., VAN DEN ESHOF, P., AND NIEROP, N. Police interviewing in france, belgium and the netherlands: something is moving. *International developments in investigative interviewing* (2009), 66–91.

[48] CORDAR, A., ROBB, A., WENDLING, A., LAMPOTANG, S., WHITE, C., AND LOK, B. Virtual role-models: Using virtual humans to train best communication practices for healthcare teams. In *INTELLIGENT VIRTUAL AGENTS, PROCEEDINGS* (2015), vol. 9238, Springer International Publishing, pp. 229–238.

[49] CORE, M. G., LANE, H. C., VAN LENT, M., GOMBOC, D., SOLOMON, S., AND ROSENBERG, M. Building explainable artificial intelligence systems. In *Proceedings of the National Conference on Artificial Intelligence* (2006), vol. 21.

[50] CRAGGS, R., AND MCGEE WOOD, M. A two dimensional annotation scheme for emotion in dialogue. In *Proceedings of AAAI Spring Symposioum on Exploring Attitude and Affect in Text: Theories and Applications* (2004).

[51] CULPEPER, J., BOUSFIELD, D., AND WICHMANN, A. Impoliteness revisited: With special reference to dynamic and prosodic aspects. *Journal of Pragmatics 35*, 10 (2003), 1545–1579.

[52] DAEL, N., MORTILLARO, M., AND SCHERER, K. R. Emotion expression in body action and posture. *Emotion 12(5)* (October 2012), 1085–1101.

[53] DAMIAN, I., BAUR, T., LUGRIN, B., GEBHARD, P., MEHLMANN, G., AND ANDRÉ, E. Games are better than books: In-situ comparison of an interactive job interview game with conventional training. In *Artificial Intelligence in Education, LNCS* (2015), vol. 9112, Springer International Publishing, pp. 84–94.

[54] DEATON, J. E., BARBA, C., SANTARELLI, T., ROSENZWEIG, L., SOUDERS, V., MCCOLLUM, C., SEIP, J., KNERR, B. W., AND SINGER, M. J. Virtual environment cultural training for operational readiness (VECTOR). *Virtual Reality 8*, 3 (jun 2005), 156–167.

[55] DEVAULT, D., MELL, J., AND GRATCH, J. Toward Natural Turn-Taking in a Virtual Human Negotiation Agent. In *AAAI Spring Symposium on Turn-taking and Coordination in Human-Machine Interaction* (2015), pp. 1–8.

[56] DIAS, J., MASCARENHAS, S., AND PAIVA, A. Fatima modular: Towards an agent architecture with a generic appraisal framework. In *Proceedings of the International Workshop on Standards for Emotion Modeling* (2011).

[57] DICE, L. Measure of the amount of ecologic association between species. *Ecology 26*, 3 (1945), 297–302.

[58] DONOHUE, W. A., AND TAYLOR, P. J. Role Effects in Negotiation: The One-Down

Phenomenon. *Negotiation Journal 23*, 3 (2007), 307–331.

[59] DOUGLAS-COWIE, E., COWIE, R., COX, C., AMIR, N., AND HEYLEN, D. The sensitive artificial listener: an induction technique for generating emotionally coloured conversation. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)* (May 2008), pp. 17–22.

[60] DU BOIS, J. W. The stance triangle. In *Stancetaking in Discourse: Subjectivity, evaluation, interaction*, R. E. (ed.), Ed. John benjamins Publishing Company, 2007, pp. 139–182.

[61] DUBOIS, D., AND PRADE, H. Fuzzy sets and probability: misunderstandings, bridges and gaps. In *Fuzzy Systems, 1993., Second IEEE International Conference on* (1993), pp. 1059–1068 vol.2.

[62] EKMAN, P., AND ROSENBERG, E. L. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, 1997.

[63] ENDRASS, B., AND ANDRÉ, E. Integration of cultural factors into the behavioural models of virtual characters. In *Natural Language Generation in Interactive Systems*, A. Stent and S. Bangalore, Eds. Cambridge University Press, 2014, pp. 227–251.

[64] EPHRATT, M. The functions of silence. *Journal of Pragmatics 40*, 11 (2008), 1909 – 1938.

[65] FILLMORE, C. J. Pragmatics and the description of discourse. *Radical Pragmatics* (1981), 143–166.

[66] GIEBELS, E. Beïnvloeding in gijzelingsonderhandelingen: De tafel van tien. *Nederlands tijdschrift voor de psychologie 57* (2002), 145–154.

[67] GIEBELS, E., AND TAYLOR, P. Interaction patterns in crisis negotiations: Persuasive arguments and cultural differences. *Journal of Applied Psychology 94* (2009), 5–19.

[68] GIFFORD, R. A lens-mapping framework for understanding the encoding and decoding of interpersonal dispositions in nonverbal behavior. *Journal of Personality and Social Psychiology 66*, 2 (1994), 398–412.

[69] GILLIES, M., AND BALLIN, D. A model of interpersonal attitude and posture generation. In *Intelligent Virtual Agents* (2003), Springer, pp. 88–92.

[70] GOFFMAN, E. *The presentation of self in everyday life*. Garden City, New York, 1959.

[71] GOFFMAN, E. *Interaction ritual*. Aldine publishing company, Chicago, 1967.

[72] GOLDBERG, J. Interrupting the discourse on interruptions An analysis in terms of relationally neutral, power- and rapport-oriented acts. *Journal of Pragmatics 14*, 6 (1990), 883–903.

[73] GRANHAG, P. A., AND HARTWIG, M. The strategic use of evidence technique. In *Detecting Deception: Current Challenges and Cognitive Approaches*, P. A. Granhag, A. Vrij, and B. Verschuere, Eds. John Wiley & Sons, Ltd, 2014, pp. 231–251.

[74] GRAVANO, A., AND HIRSCHBERG, J. A corpus-based study of interruptions in spoken dialogue. In *INTERSPEECH* (2012).

[75] GREEN JR, B. F., WOLF, A. K., CHOMSKY, C., AND LAUGHERY, K. Baseball: an automatic question-answerer. In *Papers presented at the May 9-11, 1961, western joint*

IRE-AIEE-ACM computer conference (1961), ACM, pp. 219–224.

[76] GREENBERG, C. S., BANSÉ, D., DODDINGTON, G. R., GARCIA-ROMERO, D., GODFREY, J. J., KINNUNEN, T., MARTIN, A. F., MCCREE, A., PRZYBOCKI, M., AND REYNOLDS, D. A. The nist 2014 speaker recognition i-vector machine learning challenge. In *Odyssey: The Speaker and Language Recognition Workshop* (2014).

[77] GUDJONSSON, G. H. Investigative interviewing: Recent developments and some fundamental issues. *International Review of Psychiatry 6*, 2-3 (1994), 237–245.

[78] GUDJONSSON, G. H., AND PETURSSON, H. Custodial interrogation: Why do suspects confess and how does it relate to their crime, attitude and personality? *Personality and Individual Differences 12*, 3 (1991), 295–306.

[79] GUDJONSSON, G. H., AND SINGH, K. K. The revised gudjonsson blame attribution inventory. *Personality and Individual Differences 10*, 1 (1989), 67–70.

[80] GUPTA, S., WALKER, M., AND ROMANO, D. How rude are you? Evaluating politeness and affect in interaction. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction* (2007), pp. 203–217.

[81] GURTMAN, M. B. Exploring personality with the interpersonal circumplex. *Social and Personality Psychology Compass 3*, 4 (2009), 601–619.

[82] HARTANTO, D., BRINKMAN, W.-P., KAMPMANN, I. L., MORINA, N., EMMELKAMP, P. G. M., AND NEERINCX, M. A. Design and Implementation of Home-Based Virtual Reality Exposure Therapy System with a Virtual eCoach. In *Intelligent Virtual Agents* (2015), Springer International Publishing, pp. 287–291.

[83] HARTHOLT, A., TRAUM, D., MARSELLA, S. C., SHAPIRO, A., STRATOU, G., LEUSKI, A., MORENCY, L.-P., AND GRATCH, J. All together now: Introducing the virtual human toolkit. In *Intelligent Virtual Agents* (2013).

[84] HEETER, C. Being there: The subjective experience of presence. *Presence: Teleoperators & Virtual Environments 1*, 2 (1992), 262–271.

[85] HELDNER, M., AND EDLUND, J. Pauses, gaps and overlaps in conversations. *Journal of Phonetics 38*, 4 (2010), 555–568.

[86] HESS, U., AND THIBAULT, P. Why the same expression may not mean the same when shown on different faces or seen by different people. In *Affective Information Processing*. Springer, 2009, pp. 145–158.

[87] HILL, R., GRATCH, J., MARSELLA, S., RICKEL, J., SWARTOUT, W., AND TRAUM, D. Virtual humans in the mission rehearsal exercise system. *Künstliche Intelligenz 4*, 03 (2003), 5–10.

[88] HOLMBERG, U. *Police interviews with victims and suspects of violent and sexual crimes: interviewees' experiences and interview outcomes*. PhD thesis, Department of Psychology, Stockholm University, 2004.

[89] HOLMBERG, U., AND CHRISTIANSON, S.-Å. Murderers' and sexual offenders' experiences of police interviews and their inclination to admit or deny crimes. *Behavioral Sciences & the Law 20*, 1-2 (2002), 31–45.

[90] HUANG, L., MORENCY, L.-P., AND GRATCH, J. Virtual rapport 2.0. In *Intelligent Virtual Agents* (2011), Springer, pp. 68–79.

[91] HUNG, H., AND GATICA-PEREZ, D. Identifying dominant people in meetings from

audio-visual sensors. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on* (2008), IEEE, pp. 1–6.

[92] INBAU, F. E., REID, J. E., BUCKLEY, J. P., AND JAYNE, B. C. *Criminal interrogation and confessions (5th ed.)*. Jones & Bartlett Learning, 2013.

[93] JACOBS, M. Bekennen en ontkennen van verdachten. Tech. rep., WODC, 2004.

[94] JEFFERSON, G. Glossary of transcript symbols with an introduction. In *Conversation Analysis: Studies from the first generation*, G. E. Lerner, Ed. John Benjamins, 2004, pp. 13–31.

[95] JOHN, O. P., AND SRIVASTAVA, S. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research 2*, 1999 (1999), 102–138.

[96] JONES, C. UK police interviews: A linguistic analysis of Afro-Caribbean and white British suspect interviews. *International Journal of Speech Language and the Law 15*, 2 (2008).

[97] JONES, H., AND SABOURET, N. An Affective Model for a Virtual Recruiter in a Job Interview Context. *Procedia Computer Science 15* (2012), 312–313.

[98] JONSDOTTIR, G. R., THORISSON, K. R., AND NIVEL, E. Learning smooth, human-like turntaking in realtime dialogue. In *Intelligent Virtual Agents* (2008), vol. 5208/2008, Springer Berlin / Heidelberg, pp. 162–175.

[99] KARKKAINEN, E. Stance taking in conversation: from subjectivity to intersubjectivity. *Text & Talk 26–6* (2006).

[100] KASSIN, S. M. On the psychology of confessions: Does innocence put innocents at risk? *American Psychologist 60* (2005), 215–228.

[101] KASSIN, S. M., LEO, R. A., MEISSNER, C. A., RICHMAN, K. D., COLWELL, L. H., LEACH, A.-M., AND LA FON, D. Police interviewing and interrogation: a self-report survey of police practices and beliefs. *Law and human behavior 31*, 4 (2007), 381.

[102] KASSIN, S. M., MEISSNER, C. A., AND NORWICK, R. J. "I'd know a false confession if I saw one": a comparative study of college students and police investigators. *Law and Human Behavior 29*, 2 (2005), 211–227.

[103] KENNY, P. G., HARTHOLT, A., GRATCH, J., SWARTOUT, W., TRAUM, D., MARSELLA, S. C., AND PIEPOL, D. Building interactive virtual humans for training environments. In *Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)* (Orlando, FL, Nov. 2007).

[104] KIESLER, D. J. *Contemporary Interpersonal Theory and Research, Personality, Psychopathology and Psychotherapy*. Wiley, 1996.

[105] KIILI, K. Digital game-based learning: Towards an experiential gaming model. *The Internet and higher education 8*, 1 (2005), 13–24.

[106] KLAASSEN, R., HENDRIX, J., REIDSMA, D., OP DEN AKKER, R., VAN DIJK, B., AND OP DEN AKKER, H. Elckerlyc goes mobile - Enabling natural interaction in mobile user interfaces. *International Journal on Advances in Telecommunications 6*, 1&2 (2013), 45–56.

[107] KLEINSMITH, A., AND BIANCHI-BERTHOUZE, N. Affective body expression perception and recognition: A survey. *Affective Computing, IEEE Transactions on 4*, 1 (2013), 15–

33.

[108] KOOPS, M., AND HOEVENAAR, M. Conceptual change during a serious game: Using a Lemniscate Model to compare strategies in a physics game. *Simulation & Gaming* (2012).

[109] KOPP, S., GESELLENSETTER, L., KRÄMER, N. C., AND WACHSMUTH, I. A conversational agent as museum guide–design and evaluation of a real-world application. In *Intelligent Virtual Agents* (2005), pp. 329–343.

[110] KOPP, S., VAN WELBERGEN, H., YAGHOUBZADEH, R., AND BUSCHMEIER, H. An architecture for fluid real-time conversational agents: integrating incremental output generation and input processing. *Journal on Multimodal User Interfaces 8*, 1 (nov 2014), 97–108.

[111] KRIPPENDORFF, K. *Content Analysis: An Introduction to its Methodology*. SAGE Publications, Second Edition, 2004.

[112] KRIPPENDORFF, K. Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research 30(3)* (2004), 411–433.

[113] KURZON, D. The right of silence: a socio-pragmatic model of interpretation. *Journal of Pragmatics 23*, 1 (1995), 55 – 69.

[114] LAFORGE, R., AND SUCZEK, R. F. The interpersonal dimension of personality. *Journal of Personality 24*, 1 (1955), 94–112.

[115] LAMB, M. E., STERNBERG, K. J., ORBACH, Y., HERSHKOWITZ, I., HOROWITZ, D., AND ESPLIN, P. W. The effects of intensive training and ongoing supervision on the quality of investigative interviews with alleged sex abuse victims. *Applied Develoment Science 6:3* (2002), 114–125.

[116] LEARY, T. *Interpersonal Diagnosis of Personality: Functional Theory and Methodology for Personality Evaluation*. Ronald Press, New York, 1957.

[117] LEE, K. M., AND NASS, C. Designing social presence of social actors in human computer interaction. In *Proceedings of the conference on Human factors in computing systems - CHI '03* (New York, New York, USA, 2003), ACM Press, pp. 289–296.

[118] LEUSKI, A., AND TRAUM, D. NPCEditor: Creating virtual human dialogue using information retrieval techniques. *AI Magazine 32*, 2 (2011), 42–56.

[119] LIMBRECHT-ECKLUNDT, K., SCHECK, A., JERG-BRETZKE, L., WALTER, S., HOFFMANN, H., AND TRAUE, H. The effect of forced choice on facial emotion recognition: a comparison to open verbal classification of emotion labels. *Psycho-social medicine 10* (2013), 1–8.

[120] LINSSEN, J., DE GROOT, T., THEUNE, M., AND BRUIJNES, M. Beyond simulations: Serious games for training interpersonal skills in law enforcement. In *Proceedings of the 10th annual meeting of the European Social Simulation Association (ESSA 2014)* (Barcelona, 2014), European Social Simulation Association, Universitat Autonoma de Barcelona, pp. 127–130.

[121] LINSSEN, J., THEUNE, M., DE GROOT, T., AND HEYLEN, D. Improving social awareness through thought bubbles and flashbacks of virtual characters. In *Intelligent Virtual Agents* (2015), vol. 9238, pp. 250–259.

[122] LINSSEN, J. M., THEUNE, M., AND HEYLEN, D. K. J. Taking things at face value: How

stance informs politeness of virtual agents. In *Workshop on Computers as Social Actors at IVA 2013* (2013).

[123] LOGINOV, V. Probability treatment of zadeh membership functions and their use in pattern recognition. *Eng. Cyber.* (1966), 68–69.

[124] LUCIEW, D., MULKERN, J., AND PUNAKO, R. Finding the truth: Interview and interrogation training simulations. In *The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC)* (2011).

[125] MANN, S., VRIJ, A., AND BULL, R. Detecting true lies: police officers' ability to detect suspects' lies. *Journal of Applied Psychology 89*, 1 (2004), 137.

[126] MANNING, C., AND SCHUTZE, H. *Foundations of Statistical Natural Language Processing*. The MIT Press, 1999.

[127] MARSELLA, S. C., AND GRATCH, J. Ema: A process model of appraisal dynamics. *Cognitive Systems Research 10*, 1 (2009), 70–90.

[128] MAZELAND, H. *Inleiding in de conversatieanalyse*. Uitgeverij Coutinho, 2003.

[129] MEHRABIAN, A. *Nonverbal communication*. Transaction Publishers, 2009.

[130] MEISSNER, C. A., AND KASSIN, S. M. "He's guilty!": investigator bias in judgments of truth and deception. *Law and human behavior 26*, 5 (2002), 469–480.

[131] MORENO, A., AND VAN DELDEN, R. Augmenting playing spaces to enhance the game experience: A tag game case study. *Entertainment Computing* (2016).

[132] MORINA, N., BRINKMAN, W.-P., HARTANTO, D., KAMPMANN, I. L., AND EMMELKAMP, P. M. Social interactions in virtual reality exposure therapy: A proof-of-concept pilot study. *Technology and Health Care 23*, 5 (sep 2015), 581–589.

[133] NEFF, M., WANG, Y., ABBOTT, R., AND WALKER, M. Evaluating the effect of gesture and language on personality perception in conversational agents. In *Intelligent Virtual Agents* (2010), Springer, pp. 222–235.

[134] NICOLE NOVIELLI, N., AND GENTILE, E. Modeling user interpersonal stances in affective dialogues with an eca. In *Proceedings of the Twenty-First International Conference on Software Engineering & Knowledge Engineering, SEKE* (2009), pp. 581–586.

[135] NIEROP, N. M. Het verdachtenverhoor in nederland: wat wordt verhoorders geleerd. *Nederlands Juristenblad 17* (2005), 887–890.

[136] NIEWIADOMSKI, R., BEVACQUA, E., MANCINI, M., AND PELACHAUD, C. Greta: an interactive expressive ECA system. In *Autonomous Agents and Multiagent Systems (AAMAS 2009)* (may 2009), International Foundation for Autonomous Agents and Multiagent Systems, pp. 1399–1400.

[137] NOORAEI, B., RICH, C., AND SIDNER, C. L. A Real-Time Architecture for Embodied Conversational Agents: Beyond Turn-Taking. In *7th Int. Conf. on Advances in Computer-Human Interactions* (2014), pp. 1–8.

[138] OCHS, M., SABOURET, N., AND CORRUBLE, V. Simulation of the dynamics of nonplayer characters' emotions and social relations in games. *Computational Intelligence and AI in Games, IEEE Transactions on 1*, 4 (2009), 281–297.

[139] OLSEN, D. Interview and interrogation training using a computer-simulated subject. In *The Interservice/Industry Training, Simulation & Education Conference* (1997).

[140] OP DEN AKKER, R., AND BRUIJNES, M. Computational models of social and emotional turn-taking for embodied conversational agents: a review. Tech. Rep. TR-CTIT-12-13, Centre for Telematics and Information Technology, University of Twente, Enschede, 2012.

[141] OP DEN AKKER, R., BRUIJNES, M., PETERS, R., AND KRIKKE, T. Interpersonal stance in police interviews: content analysis. *Computational Linguistics in the Netherlands Journal 3* (2013), 193–216.

[142] OP DEN AKKER, R., KLAASSEN, R., AND NIJHOLT, A. Virtual coaches for healthy lifestyle. In *Toward Robotic Socially Believable Behaving Systems - Volume II: Modeling Social Signals. Intelligent Systems Reference Library*. Springer Verlag, 2016, pp. 121–149.

[143] OP DEN AKKER, R., THEUNE, M., TRUONG, K., AND DE KOK, I. The organisation of floor in meetings and the relation with speaker addressee patterns. In *Proceedings of the 2nd International Workshop on Social Signal Processing, SSPW '10* (New York, October 2010), ACM, pp. 35–40.

[144] ORFORD, J. The rules of interpersonal complementarity: Does hostility beget hostility and dominance, submission? *Psychological Review 93*, 3 (1986), 365–377.

[145] PEREIRA, G., BRISSON, A., PRADA, R., PAIVA, A., BELLOTTI, F., KRAVCIK, M., AND KLAMMA, R. Serious games for personal and social learning & ethics: Status and trends. In *Procedia Computer Science* (2012), vol. 15, pp. 53–65.

[146] PERLIN, K. Toward interactive narrative. In *Virtual Storytelling. Using Virtual Reality Technologies for Storytelling*. Springer, 2005, pp. 135–147.

[147] PIMAN, S., AND TALIB, A. Z. An intelligent instructional tool for puppeteering in virtual shadow puppet play. In *Intelligent Technologies for Interactive Entertainment*. Springer, 2012, pp. 113–122.

[148] POESIO, M., AND ARTSTEIN, R. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the workshop on frontiers in corpus annotations ii: Pie in the sky* (2005), Association for Computational Linguistics, pp. 76–83.

[149] POGGI, I., PELACHAUD, C., DE ROSIS, F., CAROFIGLIO, V., AND DE CAROLIS, B. Greta. A Believable Embodied Conversational Agent. *Multimodal intelligent information presentation 27* (2005), 3–25.

[150] POTTER, W. J., AND LEVINE-DONNERSTEIN, D. Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research 27*, 3 (1999), 258–284.

[151] PREPIN, K., OCHS, M., AND PELACHAUD, C. Beyond backchannels: co-construction of dyadic stancce by reciprocal reinforcement of smiles between virtual agents. In *International conference CogSci (Annual conference of the cognitive science society)* (2013).

[152] RAVAJA, N., SAARI, T., TURPEINEN, M., LAARNI, J., SALMINEN, M., AND KIVIKANGAS, M. Spatial Presence and Emotions during Video Game Playing: Does It Matter with Whom You Play? *Presence: Teleoperators and Virtual Environments 15*, 4 (2006), 327–333.

[153] RAVENET, B., CAFARO, A., BIANCARDI, B., OCHS, M., AND PELACHAUD, C. Conversational behavior reflecting interpersonal attitudes in small group interactions. In *International Conference on Intelligent Virtual Agents* (2015), Springer, pp. 375–388.

[154] RAVENET, B., OCHS, M., AND PELACHAUD, C. A computational model of social attitude

effects on the nonverbal behavior for a relational agent. In *Proc. of Workshop Affect Compagnon Artificiel Interaction (WACAI 2012)* (2012).

[155] RAVENET, B., OCHS, M., AND PELACHAUD, C. From a user-created corpus of virtual agent's non-verbal behavior to a computational model of interpersonal attitudes. In *Intelligent Virtual Agents* (2013), Springer, pp. 263–274.

[156] REIDSMA, D. *Annotations and subjective machines of annotators, embodied agents, users, and other humans*. PhD thesis, Enschede, the Netherlands, October 2008.

[157] REIDSMA, D., AND CARLETTA, J. Reliability measurement without limits. *Comput. Linguist. 34*, 3 (Sept. 2008), 319–326.

[158] REIDSMA, D., AND OP DEN AKKER, R. Exploiting 'subjective' annotations. In *Proceedings of the Coling Workshop on Human Judgments in Computational Linguistics* (August 2008), R. Artstein, G. Boleda, F. Keller, and S. Schulte im Walde, Eds.

[159] REISENZEIN, R., HUDLICKA, E., DASTANI, M., GRATCH, J., HINDRIKS, K., LORINI, E., AND MEYER, J. Computational modeling of emotion: Toward improving the inter- and intradisciplinary exchange. *Affective Computing, IEEE Transactions on 4*, 3 (2013), 246–266.

[160] RICHARDSON, B. H., TAYLOR, P. J., SNOOK, B., CONCHIE, S. M., AND BENNELL, C. Language style matching and police interrogation outcomes. *Law and Human Behavior* (2014), 1–10.

[161] ROBINSON, L. F., AND REIS, H. T. The effects of interruption, gender, and status on interpersonal perceptions. *Journal of Nonverbal Behavior 13*, 3 (1989), 141–153.

[162] ROGAN, R. G. Linguistic style matching in crisis negotiations: A comparative analysis of suicidal and surrender outcomes. *Journal of Police Crisis Negotiations 11*, 1 (2011), 20–39.

[163] ROQUE, A., AND TRAUM, D. A model of compliance and emotion for potentially adversarial dialogue agents. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue* (2007), pp. 35–38.

[164] ROUCKHOUT, D., AND SCHACHT, R. Ontwikkeling van een Nederlandstalig Interpersoonlijk Circumplex. *Diagnostiekwijzer 4* (2000), 96–118.

[165] SACKS, H., SCHEGLOFF, E., AND JEFFERSON, G. A simplest systematics for the organization of turn-taking for conversation. *Language 50* (1974), 696–735.

[166] SADLER, P., ETHIER, N., GUNN, G. R., DUONG, D., AND WOODY, E. Are we on the same wavelength? interpersonal complementarity as shared cyclical patterns during interactions. *Journal of personality and social psychology 97*, 6 (2009), 1005.

[167] SCHANK, R. *Virtual learning*. McGraw-Hill New York, 1997.

[168] SCHEGLOFF, E. *Accounts of Conduct in Interaction: Interruption, Overlap and Turn-Taking*. New York: Plenum, 2000.

[169] SCHEGLOFF, E. Overlapping talk and the organization of turn-taking for conversation. *Language in Society 29*, 01 (2000), 1–63.

[170] SCHERER, K. R. What are emotions? and how can they be measured? *Social Science Information 44*, 4 (2005), 695–729.

[171] SCHRODER, M., BEVACQUA, E., COWIE, R., EYBEN, F., GUNES, H., HEYLEN, D.,

TER MAAT, M., MCKEOWN, G., PAMMI, S., PANTIC, M., ET AL. Building autonomous sensitive artificial listeners. *Affective Computing, IEEE Transactions on 3*, 2 (2012), 165–183.

[172] SCHUETZLER, R. M. *Dynamic Interviewing Agents: Effects on Deception , Nonverbal Behavior , and Social Desirability*. PhD thesis, University of Arizona, 2015.

[173] SEARLE, J. R. *Speech acts: An essay in the philosophy of language*, vol. 626. Cambridge University Press, Cambridge, 1969.

[174] SIEGEL, M. The sense-think-act paradigm revisited. In *Robotic Sensing, 2003. ROSE'03. 1st International Workshop on* (2003), IEEE, pp. 1–5.

[175] SLOETJES, H., AND WITTENBURG, P. Annotation by category–elan and iso dcr. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)* (2008).

[176] SMITH-HANEN, S. S. Effects of nonverbal behaviors on judged levels of counselor warmth and empathy. *Journal of Counseling Psychology 24(2)* (March 1977), 87–91.

[177] SNOOK, B., LUTHER, K., AND QUINLAN, H. Let 'em talk!: A field study of police questioning practices of suspects and accused persons. *Criminal Justice and Behavior 39* (October 2012), 1328–1339.

[178] SPITTERS, S., SANDERS, M., OP DEN AKKER, R., AND BRUIJNES, M. The recognition of acted interpersonal stance in police interrogations. In *4th International Conference on Cognitive Infocommunications (CogInfoCom)* (2013), IEEE, pp. 65–70.

[179] STEUNEBRINK, B. R., DASTANI, M., AND MEYER, J.-J. C. A formal model of emotion triggers: an approach for BDI agents. *Synthese 185*, 1 (2012), 83–129.

[180] STEUNEBRINK, B. R., VERGUNST, N. L., MOL, C. P., DIGNUM, F. P. M., DASTANI, M., AND MEYER, J.-J. C. A generic architecture for a companion robot. In *5th International Conference on Informatics in Control, Automation and Robotics (ICINCO)* (2008).

[181] STRÖFER, S., UFKES, E. G., BRUIJNES, M., GIEBELS, E., AND NOORDZIJ, M. L. Interviewing suspects with avatars: Avatars are more effective when perceived as human. *Frontiers in Psychology 7* (2016).

[182] STRÖMWALL, L., GRANHAG, P., AND HARTWIG, M. 10 practitioners' beliefs about deception. In *The Detection of Deception in Forensic Contexts* (2004), Cambridge University Press, pp. 229–250.

[183] STRÖMWALL, L. A., HARTWIG, M., AND GRANHAG, P. A. To act truthfully: Nonverbal behaviour and strategies during a police interrogation. *Psychology, Crime & Law 12*, 2 (2006), 207–219.

[184] SUSILO, A., EERTWEGH, V. V., DALEN, V. J., AND SCHERPBIER, A. Leary's rose to improve negotiation skills among health professionals: Experiences from a southeast asian culture. *Education for Health 26* (2013), 54–9.

[185] SVENNEVIG, J. *Getting acquainted in conversation: a study of initial interactions*. John Benjamins, Amsterdam, 1999.

[186] SWARTOUT, W. Lessons learned from virtual humans. *AI Magazine 31*, 1 (2010), 9–20.

[187] TANNEN, D. What's in a frame? Surface evidence for underlying expectations. *Framing in discourse 1456* (1993).

[188] TAYLOR, P. J. A Cylindrical Model of Communication Behavior in Crisis Negotiations. *Human Communication Research 28*, 1 (2002), 7–48.

[189] TAYLOR, P. J., JACQUES, K., GIEBELS, E., LEVINE, M., BEST, R., WINTER, J., AND ROSSI, G. Analysing forensic processes: Taking time into account. *Issues in Forensic Psychology 8* (2008), 45–57.

[190] THOMAS, K. W. Conflict and conflict management: Reflections and update. *Journal of Organizational Behavior 13*, 3 (1992), 265–274.

[191] THORISSON, K. R. Natural turn-taking needs no manual: Computational theory and model, from perception to action. In *Multimodality in Language and Speech Systems*. Kluwer Academic Publishers, 2002, pp. 173–207.

[192] THÓRISSON, K. R., GISLASON, O., JONSDOTTIR, G. R., AND THORISSON, H. T. A multiparty multimodal architecture for realtime turntaking. In *International Conference on Intelligent Virtual Agents* (2010), Springer, pp. 350–356.

[193] TICKLE-DEGNEN, L., AND ROSENTHAL, R. The nature of rapport and its nonverbal correlates. *Psychological inquiry 1*, 4 (1990), 285–293.

[194] TRAUM, D. Non-cooperative and deceptive virtual agents. *IEEE Intelligent Systems: Trends and Controversies: Computational Deception and Noncooperation 27*, 6 (2012), 66–69.

[195] TRAUM, D., MARSELLA, S. C., GRATCH, J., LEE, J., AND HARTHOLT, A. Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents. In *Intelligent Virtual Agents* (2008), pp. 117–130.

[196] TRAUM, D., ROQUE, A., LEUSKI, A., GEORGIOU, P., GERTEN, J., MARTINOVSKI, B., NARAYANAN, S., ROBINSON, S., AND VASWANI, A. Hassan: A virtual human for tactical questioning. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue* (2007), pp. 71–74.

[197] TRAUM, D., SWARTOUT, W., MARSELLA, S., AND GRATCH, J. Fight, flight, or negotiate: Believable strategies for conversing under crisis. In *Intelligent Virtual Agents* (2005), Springer, pp. 52–64.

[198] UPPER, D. The unsuccessful self-treatment of a case of "writer's block". *Journal of applied behavior analysis 7*, 3 (1974), 497–497.

[199] VAASSEN, F., AND DAELEMANS, W. Emotion classification in a serious game for training communication skills. In *Computational Linguistics in the Netherlands 2010: selected papers from the 20th CLIN meeting* (2010), LOT.

[200] VAASSEN, F., AND DAELEMANS, W. Automatic emotion classification for interpersonal communication. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis* (2011), pp. 104–110.

[201] VAASSEN, F., WAUTERS, J., VAN BROECKHOVEN, F., VAN OVERVELDT, M., DAELEMANS, W., AND ENEMAN, K. deLearyous: Training interpersonal communication skills using unconstrained text input. In *Proceedings of the 6th European Conference on Games Based Learning* (2012), pp. 505–513.

[202] VAN AMELSVOORT, A., RISPENS, I., AND GROLMAN, H. *Handleiding Verhoor*. Stapel & De Koning, 2010.

[203] VERSCHUEREN, J. *What people say and do with words*. Norwood, N.J.: Ablex, 1985.

[204] VINCIARELLI, A., PANTIC, M., AND BOURLARD, H. Social signal processing: Survey of an emerging domain. *Image and Vision Computing 27*, 12 (2009), 1743–1759.

[205] VINCIARELLI, A., PANTIC, M., HEYLEN, D. K. J., PELACHAUD, C., POGGI, I., D'ERICCO, F., AND SCHROEDER, M. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing 3*, 1 (2012), 69–87.

[206] WALKER, M. A., CAHN, J. E., AND WHITTAKER, S. J. Improvising linguistic style: Social and affective bases for agent personality. In *Proceedings of the International Conference on Autonomous Agents* (1997), pp. 96–105.

[207] WANG, W. Y., FINKELSTEIN, S., OGAN, A., BLACK, A. W., AND CASSELL, J. "Love ya, jerkface": Using sparse log-linear models to build positive (and impolite) relationships with teens. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (2012), pp. 20–29.

[208] WAPPEROM, S. Computational modelling of suspect behaviour in police interviews. Master's thesis, Human Media Interaction, University of Twente, June 2014.

[209] WAUTERS, J., VAN BROECKHOVEN, F., VAN OVERVELDT, M., ENEMAN, K., VAASSEN, F., AND DAELEMANS, W. delearyous: An interactive application for interpersonal communication training. In *Proceedings of CCIS Serious Games: The Challenge* (2011).

[210] WELLMAN, H. M., CROSS, D., AND WATSON, J. Meta-Analysis of Theory-of-Mind Development : The Truth about False Belief. *Child development 72*, 3 (2001), 655–684.

[211] WIGGINS, J. S. A psychological taxonomy of trait-descriptive terms: The interpersonal domain. *Journal of personality and social psychology 37*, 3 (1979), 395.

[212] WIGGINS, J. S. *Paradigms of Personality Assessment*. Guilford Press, 2003.

[213] WILTING, J., KRAHMER, E., AND SWERTS, M. Real vs. acted emotional speech. In *INTERSPEECH 2006: 9th International Conference on Spoken Language Processing* (2006), vol. 2, pp. 805–808.

[214] WISE, A., CHANG, J., DUFFY, T., AND DEL VALLE, R. The effects of teacher social presence on student satisfaction, engagement, and learning. *Journal of educational computing research 31*, 3 (2004), 247–271.

[215] WITMER, B. G., AND SINGER, M. J. Measuring Presence in Virtual Environments: A Presence Questionnaire. *Presence: Teleoperators and Virtual Environments 7*, 3 (1998), 225–240.

[216] WRIGHT, A. M., AND ALISON, L. Questioning sequences in canadian police interviews: Constructing and confirming the course of events? *Psychology, Crime and Law 10*, 2 (2004), 137–154.

[217] YNGVE, V. On getting a word in edgewise. In *Papers from the sixth regional meeting of the Chicago Linguistic Society* (1970), pp. 567–77.

[218] ZADEH, L. Fuzzy sets. *Information and Control 8* (1965), 333–353.

SIKS Dissertation Series

2009

1 Rasa Jurgelenaite (RUN) *Symmetric Causal Independence Models*
2 Willem Robert van Hage (VUA) *Evaluating Ontology-Alignment Techniques*
3 Hans Stol (UvT) *A Framework for Evidence-based Policy Making Using IT*
4 Josephine Nabukenya (RUN) *Improving the Quality of Organisational Policy Making using Collaboration Engineering*
5 Sietse Overbeek (RUN) *Bridging Supply and Demand for Knowledge Intensive Tasks: Based on Knowledge, Cognition, and Quality*
6 Muhammad Subianto (UU) *Understanding Classification*
7 Ronald Poppe (UT) *Discriminative Vision-Based Recovery and Recognition of Human Motion*
8 Volker Nannen (VUA) *Evolutionary Agent-Based Policy Analysis in Dynamic Environments*
9 Benjamin Kanagwa (RUN) *Design, Discovery and Construction of Service-oriented Systems*
10 Jan Wielemaker (UvA) *Logic programming for knowledge-intensive interactive applications*
11 Alexander Boer (UvA) *Legal Theory, Sources of Law & the Semantic Web*
12 Peter Massuthe (TUE, Humboldt-Universitaet zu Berlin) *Operating Guidelines for Services*
13 Steven de Jong (UM) *Fairness in Multi-Agent Systems*
14 Maksym Korotkiy (VUA) *From ontology-enabled services to service-enabled ontologies (making ontologies work in e-science with ONTO-SOA)*
15 Rinke Hoekstra (UvA) *Ontology Representation: Design Patterns and Ontologies that Make Sense*
16 Fritz Reul (UvT) *New Architectures in Computer Chess*
17 Laurens van der Maaten (UvT) *Feature Extraction from Visual Data*
18 Fabian Groffen (CWI) *Armada, An Evolving Database System*
19 Valentin Robu (CWI) *Modeling Preferences, Strategic Reasoning and Collaboration in Agent-Mediated Electronic Markets*
20 Bob van der Vecht (UU) *Adjustable Autonomy: Controling Influences on Decision Making*
21 Stijn Vanderlooy (UM) *Ranking and Reliable Classification*
22 Pavel Serdyukov (UT) *Search For Expertise: Going beyond direct evidence*
23 Peter Hofgesang (VUA) *Modelling Web Usage in a Changing Environment*
24 Annerieke Heuvelink (VUA) *Cognitive Models for Training Simulations*
25 Alex van Ballegooij (CWI) *RAM: Array Database Management through Relational Mapping*
26 Fernando Koch (UU) *An Agent-Based Model for the Development of Intelligent Mobile Services*
27 Christian Glahn (OU) *Contextual Support of social Engagement and Reflection on the Web*
28 Sander Evers (UT) *Sensor Data Management with Probabilistic Models*
29 Stanislav Pokraev (UT) *Model-Driven Semantic Integration of Service-Oriented Applications*
30 Marcin Zukowski (CWI) *Balancing vectorized query execution with bandwidth-optimized storage*
31 Sofiya Katrenko (UvA) *A Closer Look at Learning Relations from Text*
32 Rik Farenhorst (VUA) *Architectural Knowledge Management: Supporting Architects and Auditors*
33 Khiet Truong (UT) *How Does Real Affect Affect Affect Recognition In Speech?*
34 Inge van de Weerd (UU) *Advancing in Software Product Management: An Incremental Method Engineering Approach*
35 Wouter Koelewijn (UL) *Privacy en Politiegegevens: Over geautomatiseerde normatieve informatie-uitwisseling*
36 Marco Kalz (OUN) *Placement Support for Learners in Learning Networks*
37 Hendrik Drachsler (OUN) *Navigation Support for Learners in Informal Learning Networks*

38 Riina Vuorikari (OU) *Tags and self-organisation: a metadata ecology for learning resources in a multilingual context*
39 Christian Stahl (TUE, Humboldt-Universitaet zu Berlin) *Service Substitution: A Behavioral Approach Based on Petri Nets*
40 Stephan Raaijmakers (UvT) *Multinomial Language Learning: Investigations into the Geometry of Language*
41 Igor Berezhnyy (UvT) *Digital Analysis of Paintings*
42 Toine Bogers (UvT) *Recommender Systems for Social Bookmarking*
43 Virginia Nunes Leal Franqueira (UT) *Finding Multi-step Attacks in Computer Networks using Heuristic Search and Mobile Ambients*
44 Roberto Santana Tapia (UT) *Assessing Business-IT Alignment in Networked Organizations*
45 Jilles Vreeken (UU) *Making Pattern Mining Useful*
46 Loredana Afanasiev (UvA) *Querying XML: Benchmarks and Recursion*

2010

1 Matthijs van Leeuwen (UU) *Patterns that Matter*
2 Ingo Wassink (UT) *Work flows in Life Science*
3 Joost Geurts (CWI) *A Document Engineering Model and Processing Framework for Multimedia documents*
4 Olga Kulyk (UT) *Do You Know What I Know? Situational Awareness of Co-located Teams in Multidisplay Environments*
5 Claudia Hauff (UT) *Predicting the Effectiveness of Queries and Retrieval Systems*
6 Sander Bakkes (UvT) *Rapid Adaptation of Video Game AI*
7 Wim Fikkert (UT) *Gesture interaction at a Distance*
8 Krzysztof Siewicz (UL) *Towards an Improved Regulatory Framework of Free Software. Protecting user freedoms in a world of software communities and eGovernments*
9 Hugo Kielman (UL) *A Politiele gegevensverwerking en Privacy, Naar een effectieve waarborging*
10 Rebecca Ong (UL) *Mobile Communication and Protection of Children*
11 Adriaan Ter Mors (TUD) *The world according to MARP: Multi-Agent Route Planning*
12 Susan van den Braak (UU) *Sensemaking software for crime analysis*
13 Gianluigi Folino (RUN) *High Performance Data Mining using Bio-inspired techniques*
14 Sander van Splunter (VUA) *Automated Web Service Reconfiguration*
15 Lianne Bodenstaff (UT) *Managing Dependency Relations in Inter-Organizational Models*
16 Sicco Verwer (TUD) *Efficient Identification of Timed Automata, theory and practice*
17 Spyros Kotoulas (VUA) *Scalable Discovery of Networked Resources: Algorithms, Infrastructure, Applications*
18 Charlotte Gerritsen (VUA) *Caught in the Act: Investigating Crime by Agent-Based Simulation*
19 Henriette Cramer (UvA) *People's Responses to Autonomous and Adaptive Systems*
20 Ivo Swartjes (UT) *Whose Story Is It Anyway? How Improv Informs Agency and Authorship of Emergent Narrative*
21 Harold van Heerde (UT) *Privacy-aware data management by means of data degradation*
22 Michiel Hildebrand (CWI) *End-user Support for Access to Heterogeneous Linked Data*
23 Bas Steunebrink (UU) *The Logical Structure of Emotions*
24 Zulfiqar Ali Memon (VUA) *Modelling Human-Awareness for Ambient Agents: A Human Mindreading Perspective*
25 Ying Zhang (CWI) *XRPC: Efficient Distributed Query Processing on Heterogeneous XQuery Engines*
26 Marten Voulon (UL) *Automatisch contracteren*
27 Arne Koopman (UU) *Characteristic Relational Patterns*
28 Stratos Idreos (CWI) *Database Cracking: Towards Auto-tuning Database Kernels*
29 Marieke van Erp (UvT) *Accessing Natural History: Discoveries in data cleaning, structuring, and retrieval*
30 Victor de Boer (UvA) *Ontology Enrichment from Heterogeneous Sources on the Web*
31 Marcel Hiel (UvT) *An Adaptive Service Oriented Architecture: Automatically solving Interoperability Problems*
32 Robin Aly (UT) *Modeling Representation Uncertainty in Concept-Based Multimedia Retrieval*
33 Teduh Dirgahayu (UT) *Interaction Design in Service Compositions*
34 Dolf Trieschnigg (UT) *Proof of Concept: Concept-based Biomedical Information Retrieval*
35 Jose Janssen (OU) *Paving the Way for Lifelong Learning: Facilitating competence development through a learning path specification*
36 Niels Lohmann (TUe) *Correctness of services and their composition*
37 Dirk Fahland (TUe) *From Scenarios to components*
38 Ghazanfar Farooq Siddiqui (VUA) *Integrative modeling of emotions in virtual agents*
39 Mark van Assem (VUA) *Converting and Integrating Vocabularies for the Semantic Web*

40 Guillaume Chaslot (UM) *Monte-Carlo Tree Search*
41 Sybren de Kinderen (VUA) *Needs-driven service bundling in a multi-supplier setting: the computational e3-service approach*
42 Peter van Kranenburg (UU) *A Computational Approach to Content-Based Retrieval of Folk Song Melodies*
43 Pieter Bellekens (TUe) *An Approach towards Context-sensitive and User-adapted Access to Heterogeneous Data Sources, Illustrated in the Television Domain*
44 Vasilios Andrikopoulos (UvT) *A theory and model for the evolution of software services*
45 Vincent Pijpers (VUA) *e3alignment: Exploring Inter-Organizational Business-ICT Alignment*
46 Chen Li (UT) *Mining Process Model Variants: Challenges, Techniques, Examples*
47 Jahn-Takeshi Saito (UM) *Solving difficult game positions*
48 Bouke Huurnink (UvA) *Search in Audiovisual Broadcast Archives*
49 Alia Khairia Amin (CWI) *Understanding and supporting information seeking tasks in multiple sources*
50 Peter-Paul van Maanen (VUA) *Adaptive Support for Human-Computer Teams: Exploring the Use of Cognitive Models of Trust and Attention*
51 Edgar Meij (UvA) *Combining Concepts and Language Models for Information Access*

2011

1 Botond Cseke (RUN) *Variational Algorithms for Bayesian Inference in Latent Gaussian Models*
2 Nick Tinnemeier (UU) *Organizing Agent Organizations. Syntax and Operational Semantics of an Organization-Oriented Programming Language*
3 Jan Martijn van der Werf (TUe) *Compositional Design and Verification of Component-Based Information Systems*
4 Hado van Hasselt (UU) *Insights in Reinforcement Learning: Formal analysis and empirical evaluation of temporal-difference*
5 Base van der Raadt (VUA) *Enterprise Architecture Coming of Age: Increasing the Performance of an Emerging Discipline*
6 Yiwen Wang (TUe) *Semantically-Enhanced Recommendations in Cultural Heritage*
7 Yujia Cao (UT) *Multimodal Information Presentation for High Load Human Computer Interaction*
8 Nieske Vergunst (UU) *BDI-based Generation of Robust Task-Oriented Dialogues*
9 Tim de Jong (OU) *Contextualised Mobile Media for Learning*
10 Bart Bogaert (UvT) *Cloud Content Contention*
11 Dhaval Vyas (UT) *Designing for Awareness: An Experience-focused HCI Perspective*
12 Carmen Bratosin (TUe) *Grid Architecture for Distributed Process Mining*
13 Xiaoyu Mao (UvT) *Airport under Control. Multiagent Scheduling for Airport Ground Handling*
14 Milan Lovric (EUR) *Behavioral Finance and Agent-Based Artificial Markets*
15 Marijn Koolen (UvA) *The Meaning of Structure: the Value of Link Evidence for Information Retrieval*
16 Maarten Schadd (UM) *Selective Search in Games of Different Complexity*
17 Jiyin He (UvA) *Exploring Topic Structure: Coherence, Diversity and Relatedness*
18 Mark Ponsen (UM) *Strategic Decision-Making in complex games*
19 Ellen Rusman (OU) *The Mind ' s Eye on Personal Profiles*
20 Qing Gu (VUA) *Guiding service-oriented software engineering: A view-based approach*
21 Linda Terlouw (TUD) *Modularization and Specification of Service-Oriented Systems*
22 Junte Zhang (UvA) *System Evaluation of Archival Description and Access*
23 Wouter Weerkamp (UvA) *Finding People and their Utterances in Social Media*
24 Herwin van Welbergen (UT) *Behavior Generation for Interpersonal Coordination with Virtual Humans On Specifying, Scheduling and Realizing Multimodal Virtual Human Behavior*
25 Syed Waqar ul Qounain Jaffry (VUA) *Analysis and Validation of Models for Trust Dynamics*
26 Matthijs Aart Pontier (VUA) *Virtual Agents for Human Communication: Emotion Regulation and Involvement-Distance Trade-Offs in Embodied Conversational Agents and Robots*
27 Aniel Bhulai (VUA) *Dynamic website optimization through autonomous management of design patterns*
28 Rianne Kaptein (UvA) *Effective Focused Retrieval by Exploiting Query Context and Document Structure*
29 Faisal Kamiran (TUe) *Discrimination-aware Classification*
30 Egon van den Broek (UT) *Affective Signal Processing (ASP): Unraveling the mystery of emotions*
31 Ludo Waltman (EUR) *Computational and Game-Theoretic Approaches for Modeling Bounded Rationality*
32 Nees-Jan van Eck (EUR) *Methodological Advances in Bibliometric Mapping of Science*
33 Tom van der Weide (UU) *Arguing to Motivate Decisions*

34 Paolo Turrini (UU) *Strategic Reasoning in Interdependence: Logical and Game-theoretical Investigations*
35 Maaike Harbers (UU) *Explaining Agent Behavior in Virtual Training*
36 Erik van der Spek (UU) *Experiments in serious game design: a cognitive approach*
37 Adriana Burlutiu (RUN) *Machine Learning for Pairwise Data, Applications for Preference Learning and Supervised Network Inference*
38 Nyree Lemmens (UM) *Bee-inspired Distributed Optimization*
39 Joost Westra (UU) *Organizing Adaptation using Agents in Serious Games*
40 Viktor Clerc (VUA) *Architectural Knowledge Management in Global Software Development*
41 Luan Ibraimi (UT) *Cryptographically Enforced Distributed Data Access Control*
42 Michal Sindlar (UU) *Explaining Behavior through Mental State Attribution*
43 Henk van der Schuur (UU) *Process Improvement through Software Operation Knowledge*
44 Boris Reuderink (UT) *Robust Brain-Computer Interfaces*
45 Herman Stehouwer (UvT) *Statistical Language Models for Alternative Sequence Selection*
46 Beibei Hu (TUD) *Towards Contextualized Information Delivery: A Rule-based Architecture for the Domain of Mobile Police Work*
47 Azizi Bin Ab Aziz (VUA) *Exploring Computational Models for Intelligent Support of Persons with Depression*
48 Mark Ter Maat (UT) *Response Selection and Turn-taking for a Sensitive Artificial Listening Agent*
49 Andreea Niculescu (UT) *Conversational interfaces for task-oriented spoken dialogues: design aspects influencing interaction quality*

2012

1 Terry Kakeeto (UvT) *Relationship Marketing for SMEs in Uganda*
2 Muhammad Umair (VUA) *Adaptivity, emotion, and Rationality in Human and Ambient Agent Models*
3 Adam Vanya (VUA) *Supporting Architecture Evolution by Mining Software Repositories*
4 Jurriaan Souer (UU) *Development of Content Management System-based Web Applications*
5 Marijn Plomp (UU) *Maturing Interorganisational Information Systems*
6 Wolfgang Reinhardt (OU) *Awareness Support for Knowledge Workers in Research Networks*
7 Rianne van Lambalgen (VUA) *When the Going Gets Tough: Exploring Agent-based Models of Human Performance under Demanding Conditions*
8 Gerben de Vries (UvA) *Kernel Methods for Vessel Trajectories*
9 Ricardo Neisse (UT) *Trust and Privacy Management Support for Context-Aware Service Platforms*
10 David Smits (TUe) *Towards a Generic Distributed Adaptive Hypermedia Environment*
11 J. C. B. Rantham Prabhakara (TUe) *Process Mining in the Large: Preprocessing, Discovery, and Diagnostics*
12 Kees van der Sluijs (TUe) *Model Driven Design and Data Integration in Semantic Web Information Systems*
13 Suleman Shahid (UvT) *Fun and Face: Exploring non-verbal expressions of emotion during playful interactions*
14 Evgeny Knutov (TUe) *Generic Adaptation Framework for Unifying Adaptive Web-based Systems*
15 Natalie van der Wal (VUA) *Social Agents. Agent-Based Modelling of Integrated Internal and Social Dynamics of Cognitive and Affective Processes*
16 Fiemke Both (VUA) *Helping people by understanding them: Ambient Agents supporting task execution and depression treatment*
17 Amal Elgammal (UvT) *Towards a Comprehensive Framework for Business Process Compliance*
18 Eltjo Poort (VUA) *Improving Solution Architecting Practices*
19 Helen Schonenberg (TUe) *What's Next? Operational Support for Business Process Execution*
20 Ali Bahramisharif (RUN) *Covert Visual Spatial Attention, a Robust Paradigm for Brain-Computer Interfacing*
21 Roberto Cornacchia (TUD) *Querying Sparse Matrices for Information Retrieval*
22 Thijs Vis (UvT) *Intelligence, politie en veiligheidsdienst: verenigbare grootheden?*
23 Christian Muehl (UT) *Toward Affective Brain-Computer Interfaces: Exploring the Neurophysiology of Affect during Human Media Interaction*
24 Laurens van der Werff (UT) *Evaluation of Noisy Transcripts for Spoken Document Retrieval*
25 Silja Eckartz (UT) *Managing the Business Case Development in Inter-Organizational IT Projects: A Methodology and its Application*
26 Emile de Maat (UvA) *Making Sense of Legal Text*
27 Hayrettin Gurkok (UT) *Mind the Sheep! User Experience Evaluation & Brain-Computer Interface Games*
28 Nancy Pascall (UvT) *Engendering Technology Empowering Women*
29 Almer Tigelaar (UT) *Peer-to-Peer Information Retrieval*
30 Alina Pommeranz (TUD) *Designing Human-Centered Systems for Reflective Decision Making*

31 Emily Bagarukayo (RUN) *A Learning by Construction Approach for Higher Order Cognitive Skills Improvement, Building Capacity and Infrastructure*
32 Wietske Visser (TUD) *Qualitative multi-criteria preference representation and reasoning*
33 Rory Sie (OUN) *Coalitions in Cooperation Networks (COCOON)*
34 Pavol Jancura (RUN) *Evolutionary analysis in PPI networks and applications*
35 Evert Haasdijk (VUA) *Never Too Old To Learn: On-line Evolution of Controllers in Swarm- and Modular Robotics*
36 Denis Ssebugwawo (RUN) *Analysis and Evaluation of Collaborative Modeling Processes*
37 Agnes Nakakawa (RUN) *A Collaboration Process for Enterprise Architecture Creation*
38 Selmar Smit (VUA) *Parameter Tuning and Scientific Testing in Evolutionary Algorithms*
39 Hassan Fatemi (UT) *Risk-aware design of value and coordination networks*
40 Agus Gunawan (UvT) *Information Access for SMEs in Indonesia*
41 Sebastian Kelle (OU) *Game Design Patterns for Learning*
42 Dominique Verpoorten (OU) *Reflection Amplifiers in self-regulated Learning*
43 Anna Tordai (VUA) *On Combining Alignment Techniques*
44 Benedikt Kratz (UvT) *A Model and Language for Business-aware Transactions*
45 Simon Carter (UvA) *Exploration and Exploitation of Multilingual Data for Statistical Machine Translation*
46 Manos Tsagkias (UvA) *Mining Social Media: Tracking Content and Predicting Behavior*
47 Jorn Bakker (TUe) *Handling Abrupt Changes in Evolving Time-series Data*
48 Michael Kaisers (UM) *Learning against Learning: Evolutionary dynamics of reinforcement learning algorithms in strategic interactions*
49 Steven van Kervel (TUD) *Ontologogy driven Enterprise Information Systems Engineering*
50 Jeroen de Jong (TUD) *Heuristics in Dynamic Sceduling: a practical framework with a case study in elevator dispatching*

2013

1 Viorel Milea (EUR) *News Analytics for Financial Decision Support*
2 Erietta Liarou (CWI) *MonetDB/DataCell: Leveraging the Column-store Database Technology for Efficient and Scalable Stream Processing*
3 Szymon Klarman (VUA) *Reasoning with Contexts in Description Logics*
4 Chetan Yadati (TUD) *Coordinating autonomous planning and scheduling*
5 Dulce Pumareja (UT) *Groupware Requirements Evolutions Patterns*
6 Romulo Goncalves (CWI) *The Data Cyclotron: Juggling Data and Queries for a Data Warehouse Audience*
7 Giel van Lankveld (UvT) *Quantifying Individual Player Differences*
8 Robbert-Jan Merk (VUA) *Making enemies: cognitive modeling for opponent agents in fighter pilot simulators*
9 Fabio Gori (RUN) *Metagenomic Data Analysis: Computational Methods and Applications*
10 Jeewanie Jayasinghe Arachchige (UvT) *A Unified Modeling Framework for Service Design*
11 Evangelos Pournaras (TUD) *Multi-level Reconfigurable Self-organization in Overlay Services*
12 Marian Razavian (VUA) *Knowledge-driven Migration to Services*
13 Mohammad Safiri (UT) *Service Tailoring: User-centric creation of integrated IT-based homecare services to support independent living of elderly*
14 Jafar Tanha (UvA) *Ensemble Approaches to Semi-Supervised Learning Learning*
15 Daniel Hennes (UM) *Multiagent Learning: Dynamic Games and Applications*
16 Eric Kok (UU) *Exploring the practical benefits of argumentation in multi-agent deliberation*
17 Koen Kok (VUA) *The PowerMatcher: Smart Coordination for the Smart Electricity Grid*
18 Jeroen Janssens (UvT) *Outlier Selection and One-Class Classification*
19 Renze Steenhuizen (TUD) *Coordinated Multi-Agent Planning and Scheduling*
20 Katja Hofmann (UvA) *Fast and Reliable Online Learning to Rank for Information Retrieval*
21 Sander Wubben (UvT) *Text-to-text generation by monolingual machine translation*
22 Tom Claassen (RUN) *Causal Discovery and Logic*
23 Patricio de Alencar Silva (UvT) *Value Activity Monitoring*
24 Haitham Bou Ammar (UM) *Automated Transfer in Reinforcement Learning*
25 Agnieszka Anna Latoszek-Berendsen (UM) *Intention-based Decision Support. A new way of representing and implementing clinical guidelines in a Decision Support System*
26 Alireza Zarghami (UT) *Architectural Support for Dynamic Homecare Service Provisioning*
27 Mohammad Huq (UT) *Inference-based Framework Managing Data Provenance*
28 Frans van der Sluis (UT) *When Complexity becomes Interesting: An Inquiry into the Information eXperience*
29 Iwan de Kok (UT) *Listening Heads*
30 Joyce Nakatumba (TUe) *Resource-Aware Business Process Management: Analysis and Support*
31 Dinh Khoa Nguyen (UvT) *Blueprint Model and Language for Engineering Cloud Applications*

32 Kamakshi Rajagopal (OUN) *Networking For Learning: The role of Networking in a Lifelong Learner's Professional Development*
33 Qi Gao (TUD) *User Modeling and Personalization in the Microblogging Sphere*
34 Kien Tjin-Kam-Jet (UT) *Distributed Deep Web Search*
35 Abdallah El Ali (UvA) *Minimal Mobile Human Computer Interaction*
36 Than Lam Hoang (TUe) *Pattern Mining in Data Streams*
37 Dirk Börner (OUN) *Ambient Learning Displays*
38 Eelco den Heijer (VUA) *Autonomous Evolutionary Art*
39 Joop de Jong (TUD) *A Method for Enterprise Ontology based Design of Enterprise Information Systems*
40 Pim Nijssen (UM) *Monte-Carlo Tree Search for Multi-Player Games*
41 Jochem Liem (UvA) *Supporting the Conceptual Modelling of Dynamic Systems: A Knowledge Engineering Perspective on Qualitative Reasoning*
42 Léon Planken (TUD) *Algorithms for Simple Temporal Reasoning*
43 Marc Bron (UvA) *Exploration and Contextualization through Interaction and Concepts*

2014

1 Nicola Barile (UU) *Studies in Learning Monotone Models from Data*
2 Fiona Tuliyano (RUN) *Combining System Dynamics with a Domain Modeling Method*
3 Sergio Raul Duarte Torres (UT) *Information Retrieval for Children: Search Behavior and Solutions*
4 Hanna Jochmann-Mannak (UT) *Websites for children: search strategies and interface design - Three studies on children's search performance and evaluation*
5 Jurriaan van Reijsen (UU) *Knowledge Perspectives on Advancing Dynamic Capability*
6 Damian Tamburri (VUA) *Supporting Networked Software Development*
7 Arya Adriansyah (TUe) *Aligning Observed and Modeled Behavior*
8 Samur Araujo (TUD) *Data Integration over Distributed and Heterogeneous Data Endpoints*
9 Philip Jackson (UvT) *Toward Human-Level Artificial Intelligence: Representation and Computation of Meaning in Natural Language*
10 Ivan Salvador Razo Zapata (VUA) *Service Value Networks*
11 Janneke van der Zwaan (TUD) *An Empathic Virtual Buddy for Social Support*
12 Willem van Willigen (VUA) *Look Ma, No Hands: Aspects of Autonomous Vehicle Control*
13 Arlette van Wissen (VUA) *Agent-Based Support for Behavior Change: Models and Applications in Health and Safety Domains*
14 Yangyang Shi (TUD) *Language Models With Meta-information*
15 Natalya Mogles (VUA) *Agent-Based Analysis and Support of Human Functioning in Complex Socio-Technical Systems: Applications in Safety and Healthcare*
16 Krystyna Milian (VUA) *Supporting trial recruitment and design by automatically interpreting eligibility criteria*
17 Kathrin Dentler (VUA) *Computing healthcare quality indicators automatically: Secondary Use of Patient Data and Semantic Interoperability*
18 Mattijs Ghijsen (UvA) *Methods and Models for the Design and Study of Dynamic Agent Organizations*
19 Vinicius Ramos (TUe) *Adaptive Hypermedia Courses: Qualitative and Quantitative Evaluation and Tool Support*
20 Mena Habib (UT) *Named Entity Extraction and Disambiguation for Informal Text: The Missing Link*
21 Kassidy Clark (TUD) *Negotiation and Monitoring in Open Environments*
22 Marieke Peeters (UU) *Personalized Educational Games: Developing agent-supported scenario-based training*
23 Eleftherios Sidirourgos (UvA/CWI) *Space Efficient Indexes for the Big Data Era*
24 Davide Ceolin (VUA) *Trusting Semi-structured Web Data*
25 Martijn Lappenschaar (RUN) *New network models for the analysis of disease interaction*
26 Tim Baarslag (TUD) *What to Bid and When to Stop*
27 Rui Jorge Almeida (EUR) *Conditional Density Models Integrating Fuzzy and Probabilistic Representations of Uncertainty*
28 Anna Chmielowiec (VUA) *Decentralized k-Clique Matching*
29 Jaap Kabbedijk (UU) *Variability in Multi-Tenant Enterprise Software*
30 Peter de Cock (UvT) *Anticipating Criminal Behaviour*
31 Leo van Moergestel (UU) *Agent Technology in Agile Multiparallel Manufacturing and Product Support*
32 Naser Ayat (UvA) *On Entity Resolution in Probabilistic Data*
33 Tesfa Tegegne (RUN) *Service Discovery in eHealth*
34 Christina Manteli (VUA) *The Effect of Governance in Global Software Development: Analyzing Transactive Memory Systems*
35 Joost van Ooijen (UU) *Cognitive Agents in Virtual Worlds: A Middleware Design Approach*

36 Joos Buijs (TUe) *Flexible Evolutionary Algorithms for Mining Structured Process Models*
37 Maral Dadvar (UT) *Experts and Machines United Against Cyberbullying*
38 Danny Plass-Oude Bos (UT) *Making brain-computer interfaces better: improving usability through post-processing*
39 Jasmina Maric (UvT) *Web Communities, Immigration, and Social Capital*
40 Walter Omona (RUN) *A Framework for Knowledge Management Using ICT in Higher Education*
41 Frederic Hogenboom (EUR) *Automated Detection of Financial Events in News Text*
42 Carsten Eijckhof (CWI/TUD) *Contextual Multidimensional Relevance Models*
43 Kevin Vlaanderen (UU) *Supporting Process Improvement using Method Increments*
44 Paulien Meesters (UvT) *Intelligent Blauw: Intelligence-gestuurde politiezorg in gebiedsgebonden eenheden*
45 Birgit Schmitz (OUN) *Mobile Games for Learning: A Pattern-Based Approach*
46 Ke Tao (TUD) *Social Web Data Analytics: Relevance, Redundancy, Diversity*
47 Shangsong Liang (UvA) *Fusion and Diversification in Information Retrieval*

2015

1 Niels Netten (UvA) *Machine Learning for Relevance of Information in Crisis Response*
2 Faiza Bukhsh (UvT) *Smart auditing: Innovative Compliance Checking in Customs Controls*
3 Twan van Laarhoven (RUN) *Machine learning for network data*
4 Howard Spoelstra (OUN) *Collaborations in Open Learning Environments*
5 Christoph Bösch (UT) *Cryptographically Enforced Search Pattern Hiding*
6 Farideh Heidari (TUD) *Business Process Quality Computation: Computing Non-Functional Requirements to Improve Business Processes*
7 Maria-Hendrike Peetz (UvA) *Time-Aware Online Reputation Analysis*
8 Jie Jiang (TUD) *Organizational Compliance: An agent-based model for designing and evaluating organizational interactions*
9 Randy Klaassen (UT) *HCI Perspectives on Behavior Change Support Systems*
10 Henry Hermans (OUN) *OpenU: design of an integrated system to support lifelong learning*
11 Yongming Luo (TUe) *Designing algorithms for big graph datasets: A study of computing bisimulation and joins*
12 Julie M. Birkholz (VUA) *Modi Operandi of Social Network Dynamics: The Effect of Context on Scientific Collaboration Networks*
13 Giuseppe Procaccianti (VUA) *Energy-Efficient Software*
14 Bart van Straalen (UT) *A cognitive approach to modeling bad news conversations*
15 Klaas Andries de Graaf (VUA) *Ontology-based Software Architecture Documentation*
16 Changyun Wei (UT) *Cognitive Coordination for Cooperative Multi-Robot Teamwork*
17 André van Cleeff (UT) *Physical and Digital Security Mechanisms: Properties, Combinations and Trade-offs*
18 Holger Pirk (CWI) *Waste Not, Want Not!: Managing Relational Data in Asymmetric Memories*
19 Bernardo Tabuenca (OUN) *Ubiquitous Technology for Lifelong Learners*
20 Loïs Vanhée (UU) *Using Culture and Values to Support Flexible Coordination*
21 Sibren Fetter (OUN) *Using Peer-Support to Expand and Stabilize Online Learning*
22 Zhemin Zhu (UT) *Co-occurrence Rate Networks*
23 Luit Gazendam (VUA) *Cataloguer Support in Cultural Heritage*
24 Richard Berendsen (UvA) *Finding People, Papers, and Posts: Vertical Search Algorithms and Evaluation*
25 Steven Woudenberg (UU) *Bayesian Tools for Early Disease Detection*
26 Alexander Hogenboom (EUR) *Sentiment Analysis of Text Guided by Semantics and Structure*
27 Sándor Héman (CWI) *Updating compressed column-stores*
28 Janet Bagorogoza (TiU) *Knowledge Management and High Performance: The Uganda Financial Institutions Model for HPO*
29 Hendrik Baier (UM) *Monte-Carlo Tree Search Enhancements for One-Player and Two-Player Domains*
30 Kiavash Bahreini (OUN) *Real-time Multimodal Emotion Recognition in E-Learning*
31 Yakup Koç (TUD) *On Robustness of Power Grids*
32 Jerome Gard (UL) *Corporate Venture Management in SMEs*
33 Frederik Schadd (UM) *Ontology Mapping with Auxiliary Resources*
34 Victor de Graaff (UT) *Geosocial Recommender Systems*
35 Junchao Xu (TUD) *Affective Body Language of Humanoid Robots: Perception and Effects in Human Robot Interaction*

2016

1 Syed Saiden Abbas (RUN) *Recognition of Shapes by Humans and Machines*
2 Michiel Christiaan Meulendijk (UU) *Optimizing medication reviews through decision support: prescribing a better pill to swallow*
3 Maya Sappelli (RUN) *Knowledge Work in Context: User Centered Knowledge Worker Support*

4 Laurens Rietveld (VUA) *Publishing and Consuming Linked Data*
5 Evgeny Sherkhonov (UvA) *Expanded Acyclic Queries: Containment and an Application in Explaining Missing Answers*
6 Michel Wilson (TUD) *Robust scheduling in an uncertain environment*
7 Jeroen de Man (VUA) *Measuring and modeling negative emotions for virtual training*
8 Matje van de Camp (TiU) *A Link to the Past: Constructing Historical Social Networks from Unstructured Data*
9 Archana Nottamkandath (VUA) *Trusting Crowdsourced Information on Cultural Artefacts*
10 George Karafotias (VUA) *Parameter Control for Evolutionary Algorithms*
11 Anne Schuth (UvA) *Search Engines that Learn from Their Users*
12 Max Knobbout (UU) *Logics for Modelling and Verifying Normative Multi-Agent Systems*
13 Nana Baah Gyan (VUA) *The Web, Speech Technologies and Rural Development in West Africa: An ICT4D Approach*
14 Ravi Khadka (UU) *Revisiting Legacy Software System Modernization*
15 Steffen Michels (RUN) *Hybrid Probabilistic Logics: Theoretical Aspects, Algorithms and Experiments*
16 Guangliang Li (UvA) *Socially Intelligent Autonomous Agents that Learn from Human Reward*
17 Berend Weel (VUA) *Towards Embodied Evolution of Robot Organisms*
18 Albert Meroño Peñuela (VUA) *Refining Statistical Data on the Web*
19 Julia Efremova (TUe) *Mining Social Structures from Genealogical Data*
20 Daan Odijk (UvA) *Context & Semantics in News & Web Search*
21 Alejandro Moreno Célleri (UT) *From Traditional to Interactive Playspaces: Automatic Analysis of Player Behavior in the Interactive Tag Playground*
22 Grace Lewis (VUA) *Software Architecture Strategies for Cyber-Foraging Systems*
23 Fei Cai (UvA) *Query Auto Completion in Information Retrieval*
24 Brend Wanders (UT) *Repurposing and Probabilistic Integration of Data; An Iterative and data model independent approach*
25 Julia Kiseleva (TUe) *Using Contextual Information to Understand Searching and Browsing Behavior*
26 Dilhan Thilakarathne (VUA) *In or Out of Control: Exploring Computational Models to Study the Role of Human Awareness and Control in Behavioural Choices, with Applications in Aviation and Energy Management Domains*
27 Wen Li (TUD) *Understanding Geo-spatial Information on Social Media*
28 Mingxin Zhang (TUD) *Large-scale Agent-based Social Simulation - A study on epidemic prediction and control*
29 Nicolas Höning (TUD) *Peak reduction in decentralised electricity systems -Markets and prices for flexible planning*
30 Ruud Mattheij (UvT) *The Eyes Have It*
31 Mohammad Khelghati (UT) *Deep web content monitoring*
32 Eelco Vriezekolk (UT) *Assessing Telecommunication Service Availability Risks for Crisis Organisations*
33 Peter Bloem (UvA) *Single Sample Statistics, exercises in learning from just one example*
34 Dennis Schunselaar (TUe) *Configurable Process Trees: Elicitation, Analysis, and Enactment*
35 Zhaochun Ren (UvA) *Monitoring Social Media: Summarization, Classification and Recommendation*
36 Daphne Karreman (UT) *Beyond R2D2: The design of nonverbal interaction behavior optimized for robot-specific morphologies*
37 Giovanni Sileno (UvA) *Aligning Law and Action - a conceptual and computational inquiry*
38 Andrea Minuto (UT) *MATERIALS THAT MATTER - Smart Materials meet Art & Interaction Design*

CTIT

CTIT PH.D. THESIS SERIES NO. 16-408
ISSN: 1381-3617

ISBN 978-90-365-4203-6