

Lecture Notes in Artificial Intelligence 1814

Subseries of Lecture Notes in Computer Science

Ana Paiva (Ed.)

Affective Interactions

**Towards a New Generation
of Computer Interfaces**



Springer

Lecture Notes in Artificial Intelligence 1814

Subseries of Lecture Notes in Computer Science

Edited by J. G. Carbonell and J. Siekmann

Lecture Notes in Computer Science

Edited by G.Goos, J. Hartmanis, and J. van Leeuwen

Berlin
Heidelberg
New York
Barcelona
Hong Kong
London
Milan
Paris
Singapore
Tokyo

Ana Paiva (Ed.)

Affective Interactions

Towards a New Generation
of Computer Interfaces

Series Editors

Jaime G. Carbonell, Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editor

Ana Paiva
Instituto Superior Técnico
Departamento de Engenharia Informática, INESC
Rua Alves Redol, 9, 1000 Lisboa, Portugal
E-mail: ana.paiva@inesc.pt

Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Affective interactions : towards a new generation of computer
interfaces / Ana Paiva (ed.). - Berlin ; Heidelberg ; New York ;
Barcelona ; Hong Kong ; London ; Milan ; Paris ; Singapore ; Tokyo :
Springer, 2000
(Lecture notes in computer science ; Vol. 1814 : Lecture notes in
artificial intelligence)
ISBN 3-540-41520-3

CR Subject Classification (1998): H.5, I.3, I.2

ISBN 3-540-41520-3 Springer-Verlag Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer-Verlag. Violations are liable for prosecution under the German Copyright Law.

Springer is a company in the BertelsmannSpringer publishing group.
© Springer-Verlag Berlin Heidelberg 2000
Printed in Germany

Typesetting: Camera-ready by author, data conversion by DA-TeX Gerd Blumenstein
Printed on acid-free paper SPIN: 10720296 06/3142 5 4 3 2 1 0

Preface

Affective computing is a fascinating new area of research emerging in computer science. It dwells on problems where “computing is related to, arises from or deliberately influences emotions” (*Affective Computing* by Picard 1997). Following this new research direction and considering the human element as crucial in designing and implementing interactive intelligent interfaces, affective computing is now influencing the way we shape, design, construct and evaluate human/computer and computer-mediated communication. But how can computers play a role in affective interactions? How can they induce emotions, recognize emotions, represent or express emotions? What kinds of interactions and what kinds of mechanisms should they have in order for affective interactions to be established with users?

To discuss these questions we decided to organise a one and a half day workshop entitled “Affective interactions: towards a new generation of interfaces” in Sienna in conjunction with the 1999 I3 Annual Conference. This book brings together some selected papers presented as first drafts at the workshop plus some extra contributions such as an interview with Prof. Rosalind Picard, a pioneer researcher in this field.

The papers combine different perspectives, theories and cases showing how to bring an affective dimension into the interaction between users and computer applications.

I would like to take this opportunity to thank many people who were involved in the preparation of this book, in particular the programme committee of the workshop and the reviewers of the book. I also want to thank all my colleagues at INESC-GAIVA group in Portugal and at the Imperial College, Department of Electrical and Electronic Engineering, London, UK. Special thanks to Carlos Martinho, Marco Vala and Nikos Drakos who helped me in the preparation of the original proceedings used during the workshop and in the finalization of this volume. Thanks to the I3 conference organisers for providing such an inspiring setting to host the event. Thanks also to Mr. Alfred Hoffmann from Springer for his help and especially for his patience during the preparation of this volume.

It is my hope that the research presented in this book can contribute greatly to an increased understanding of the role that affect plays in human/computer and computer-mediated interactions as well as for stimulating further work in this challenging new area of research.

Programme Committee of the IWAI

- Workshop Organiser: Ana Paiva, INESC and IST, Portugal
- Programme Committee: Elisabeth André, DFKI GmbH, Germany
Yasmine Arafa, Imperial College, UK
Gene Ball, Microsoft Research, USA
Luis Botelho, ADETTI, Portugal
Dolores Cañamero, IIA-CSIC, Spain
Fiorella De Rosis, University of Bari, Italy
Kristina Höök, SICS, Sweden
Abe Mamdani, Imperial College, UK
Carlos Martinho, INESC and IST, Portugal
Ana Paiva, INESC and IST, Portugal
Daniela Petrelli, ITC-IRST, Italy
Paolo Petta, ÖFAI, Austria
Juan Velásquez, MIT AI Laboratory, USA
Annika Waern, SICS, Sweden
- Reviewers: Elisabeth André, DFKI GmbH, Germany
Yasmine Arafa, Imperial College, UK
Gene Ball, Microsoft Research, USA
Luis Botelho, ADETTI, Portugal
Dolores Cañamero, IIA-CSIC, Spain
Fiorella De Rosis, University of Bari, Italy
Kristina Höök, SICS, Sweden
Isabel Machado, INESC, Portugal
Abe Mamdani, Imperial College, UK
Nuno Mamede, IST, Portugal
Carlos Martinho, INESC and IST, Portugal
Ana Paiva, INESC and IST, Portugal
Daniela Petrelli, ITC-IRST, Italy
Paolo Petta, ÖFAI, Austria
Phoebe Sengers, GMD, Germany
Juan Velásquez, MIT AI Laboratory, USA
Rodrigo Ventura, IST, Portugal
Annika Waern, SICS, Sweden

Sponsoring Institutions

- i3 Net – The European Network of Excellence for Intelligent Information Interfaces (<http://www.i3net.org/>)
AgentLink – The European Network of Excellence for Agent-Based Computing (<http://www.AgentLink.org/>)

Table of Contents

Affective Interactions: Toward a New Generation of Computer Interfaces? ... 1	
<i>Ana Paiva</i>	
Listen to Your Heart Rate: Counting the Cost of Media Quality 9	
<i>Gillian M. Wilson and M. Angela Sasse</i>	
Effective Affective in Intelligent Systems – Building on Evidence of Empathy in Teaching and Learning 21	
<i>Bridget Cooper, Paul Brna and Alex Martins</i>	
The Communication of Meaningful Emotional Information for Children Interacting with Virtual Actors 35	
<i>Pat George and Malcolm McIlhagga</i>	
Emotion and Facial Expression 49	
<i>Thomas Wehrle and Susanne Kaiser</i>	
A Cognitive Approach to Affective User Modeling 64	
<i>Carlos Martinho, Isabel Machado and Ana Paiva</i>	
Affective Appraisal versus Cognitive Evaluation in Social Emotions and Interactions 76	
<i>Cristiano Castelfranchi</i>	
An Emotion-Based “Conscious” Software Agent Architecture 107	
<i>Lee McCauley, Stan Franklin and Myles Bogner</i>	
Redesigning the Agents’ Decision Machinery 121	
<i>Luis Antunes and Helder Coelho</i>	
Artificial Emotion and Emotion Learning: Emotions as Value Judgements 138	
<i>Stevo Bozinovski</i>	
Integrating Models of Personality and Emotions into Lifelike Characters 150	
<i>Elisabeth André, Martin Klesen, Patrick Gebhard, Steve Allen and Thomas Rist</i>	
Why Should Agents Be Emotional for Entertaining Users? A Critical Analysis 166	
<i>Paola Rizzo</i>	
Emotional Meaning and Expression in Animated Faces 182	
<i>Isabella Poggi and Catherine Pelachaud</i>	
Relating Personality and Behavior: Posture and Gestures 196	
<i>Gene Ball and Jack Breese</i>	

Affective Natural Language Generation 204
Fiorella de Rosis and Floriana Grasso

An Interview with Rosalind Picard, Author of “Affective Computing” 219
Rosalind Picard

Subject Index 229

Author Index 235

Affective Interactions: Toward a New Generation of Computer Interfaces?

Ana Paiva

IST and Instituto de Engenharia de Sistemas e Computadores
Rua Alves Redol, 9, 1000 Lisboa, Portugal.
Ana.Paiva@inesc.pt

Although emotions were, for a long time, considered undesirable for rational behaviour, there is now evidence in Neuroscience and Psychology that shows emotions to be an important factor in problem solving capabilities and intelligence in general [8]. As a result, a strong new field is emerging in computer science: Affective Computing, i.e. “computing that relates to, arises from or deliberately influences emotions” [21].

Following this new research orientation and considering the human-element as crucial in designing and implementing interactive intelligent interfaces, this book provides a view on the impact of affect in such interfaces. Thus, the title “Affect in Interactions”, may be misleading, since we use “affective interactions” not only to refer to human-human interactions, but mostly, as a reference to computer mediated, generated or interpreted affective interactions. That is, we bring computers into the affective loop by exploring affect in the human/computer interaction loop. Secondly, we used the word “affective” (or affective phenomena) rather than “emotional”, since although frequently used indistinguishably, the term “affective” can be seen as more wide-ranging and therefore include, together with emotions, other states such as moods or sentiments (see [11] and [12] for good discussions on the differences between several affective phenomena).

Finally, assuming that interaction with computers should be grounded on how people interact with each other, and considering that emotion plays a fundamental role in such communication, it is not only natural but indispensable to study affect in human/computer interactions. In doing so, we expect to find techniques and approaches that will allow us to augment the actual communication channel between machines and humans. But how can machines play a role in affective communication? How can they induce emotion, recognize emotion or represent emotions? What kinds of interactions and what kinds of mechanisms should they have in order for affective interactions to be established with users? This book tries, if not to provide some answers, at least to motivate for the challenging problems of finding such answers.

1 Do We Really Need to Enrich Human/Computer Interactions with an Affective Dimension?

One can look at today’s computer applications as a good starting point towards the understanding of the role that affective elements play in the relations between

humans and machines. But to do that, evidences must be gathered on how current Human/Machine interactions influence the user's affective states.

Following such line of research, G. Wilson and A. Sasse [29] present an interesting result of measuring stress levels of users interacting in a multimedia conferencing environment with audio and video communication. In their research they found that, although users did not register subjectively the difference between receiving 5 or 25 video frames per second (only 16% reported a frame rate change) their physiological signals reported a significant increase of their stress level (75 % of the participants showed significant increase in their stress level). These results show that more research needs to be done on the analysis of affect elements present in current human/computer interactions.

Although it is clear that affective interactions are richer and allow stronger human relations, what is not so clear is in what situations one should call for such type of interactions with computers. For example, if we want to buy and sell shares on the Web, we may want a cold not emotional interaction (see R. Picard interview in this book [22]). On the other hand, if we want an environment that advises mothers in critical and stressfull situations [17], we will perhaps want to consider the actual emotional state of the mother and take that into account for the advice to give. So, evidences must be gathered to understand in which situations it is effective to explore affective interactions.

In this book, Cooper et. al. [7] present an interesting study of the role of empathy in teacher-children interactions. Based on evidence collected with children and teachers, the paper provides a good example of interactions where affect plays a fundamental role in achieving effective learning interactions. In the paper some characteristics on what is considered to be an "emphatic" teacher are provided, such as for example, being open, warm, relaxed, smiling frequently, using body language, among others. This analysis must serve as a reference for those who intend to build computer based learning environments that provide empathic interactions with the learners.

2 Capturing the User's Affective States

Humans establish affective relations with the environment from birth. Starting with the strong relation between the newborn baby and her mother, through the relations children establish with their transitional objects, such as a teddy bear or blanket, affect plays an important role in human development. Babies at 4 weeks give joy to their parents with their first smiles and at three months infants can show an expression of rage denoting disappointment [27]. But infants not only express emotions, but are also able to recognize emotions. Infants as young as five months can discriminate among facial and vocal expressions of happy, sad and angry emotions (see [26]). Voice and face become the channels for their affective interactions.

For us humans, recognizing that the person seating in front of us in the train gets distressed when the train stops for a while, with no apparent good reason, seems to be not too difficult, and we do it all the time (with more or less

success). The person may look repetitively at the watch, start shuffling his feet, or simply smile in a forced manner. Some of these signals may reveal distress or anxiety. For machines, recognizing emotions and distinguishing them from other communicative or social signals is a difficult task since it needs to be seen within a temporal and situational context [28]. But the advances in emotion recognition seem to have gone closer to science fiction predictions than ever before. Some of the most impressive results on emotion recognition were achieved by the group of Rosalind Picard at the MIT media lab (see interview with Rosalind Picard in this book [22]). The recognition of 8 emotional states (neutral; anger; hate; grief; platonic love; romantic love; joy and reverence) was done with a high success rate of 80%. Such rate was attained in a controlled situation with one subject (actress) expressing a sequence of emotions. Although, extremely impressive, there is still a long way to go for machines to be able to capture emotions by different people, in different days, in non-controlled environments. But it is a very good start.

But, as emotions can be expressed in different ways, through bodily responses, to language and behaviour changes, they therefore can be captured also using different modalities and approaches, such as: recognition of emotions in speech and voice intonation; recognition of emotions in facial expressions; recognition of emotions in body gestures, posture or movement; physiological signals recognition (such as respiration, heart rate, pulse, temperature, galvanic skin conductivity, etc); and situation evaluation.

So, it is not surprising that some of the techniques for affect acquisition rely on the physiological signals, capturing the body responses of the user, and others rely more on the interpretation of the situation and thus evaluating the cognitive response of the user. In this book some of these different approaches for emotion recognition are discussed. Although they are different, as they capture different expressions of emotion, they can be seen as complementary and part of a large multi-modal affective sensory system. But, as argued in [21], the best recognition is likely to come from the combination of the different modalities and including not only low-level signal recognition but also higher-level reasoning about the situation.

Wehrle and Kaiser [28] start with an argument for automatically interpreting gestures and facial expressions in order to improve human-machine interactions. In their paper, they show how to use an appraisal based approach to understand the relation between emotion and facial expressions. Using FACS [11], they established a link between facial actions with appraisal checks. Their results were obtained using the Geneva Appraisal Manipulation Environment (GAME), which is a tool for generating experimental computer games that translate psychological postulates into specific micro-world scenarios. While playing the game, participants are videotaped and these tape recordings allow an automatic analysis of the participants facial behaviour. The results obtained by Wehrle and Kaiser [28] Kaiser stress the advantages of adopting an appraisal-based approach which combines the subjective evaluation of a situation with the outcome of the appraisal process to other components of emotions.

Using a different modality G. Wilson and A. Sasse [29] show how measuring physiological signals of negative emotions (focusing on stress) can be seen as a more objective form of obtaining the user’s cost of media presentations. The measurement of the stress level was done by analyzing the Blood Volume Pressure (BVP), the Heart Rate (HR) and Galvanic Skin Response (GSR). These signals were chosen as they are non-obtrusive and at the same time good indicators of stress levels. The capture is done with equipment ProComp and the signal analysis was based on the fact that high heart rate is associated with an anxious state, GSR increase is also a good indicator of stress and decrease in BVP amplitude suggests that a person is fearful and anxious.

Finally, the work by Paiva, Martinho & Machado (see [18]) shows a way by which applications can take advantage of their current user modeling techniques to obtain some situation modeling needed for a cognitive approach to affective user modeling. Following a similar line of research as [28], in the presented work the OCC [20] theory of emotions is used to illustrate how to capture elements such as preferences or goals of the user, and use them to perform inferences about the affective state of the user. Such approach is then contextualized in a collaborative story creation environment called *Teatrix* [16].

3 Representing and Reasoning with Affective States

In order for affective interactions to be established between machines and humans, computers may not only recognise the user’s affective states but also, interpret and represent such affective states. Further, to exhibit affective behaviour in a coherent manner, computers themselves may incorporate models of emotions through what R. Picard calls “emotion synthesis”. Indeed, over the last few years we have seen a growing interest on emotion theories and architectures aiming at obtaining “machine emotions”. However, this growth does not mean that there is a clear consensus between the research communities on what emotions are and what theories are better applied to obtain computer emotions.

For example, Castelfranchi ([6] in this book) sees emotions as complex states constituted by the integration of mental, somatic, cognitive and motivational components. The basic constituent elements of emotions are beliefs, evaluations, goals, arousal- i. e. somatic, and the “tendency towards action”. In the paper Castelfranchi looks at two aspects of emotions: (1) that they constitute rich mental states and (2) that they are “felt”. This leads to a distinction between “cognitive evaluation” versus “intuitive and implicit affective appraisal”. Castelfranchi argues that a purely functional explanation of emotions (often found in some theories of emotion) is incomplete as it ignores their subjective face. This does not mean that beliefs, goals and expectations, elements associated with the cognitive view of emotion, are second class elements in emotion generation. In fact, such elements are crucial for what Castelfranchi considers as reason-based evaluation or “cognitive evaluation”. This analysis of “cognitive evaluation” versus “intuitive and implicit affective appraisal” is extremely important as it provides

us a framework for establishing a link between two different ways of addressing emotion synthesis.

Following the cognitive evaluation view discussed by Castelfranchi, Antunes & Coelho (see [3]) consider that one of the fundamental elements in such evaluation are "values". They propose an architecture, which includes values as the mechanism for goals adoption. Although relying deeply on a strong AI position that sees emotions only at a cognitive level, the advantage of this architecture is that there is a clear path between goals adoption and values, and thus, between goals adoption and cognitive evaluation that leads to emotions.

In McCauley, Franklin and Bogner's paper [19] the role of emotions appears in an integrated architecture that focuses on achieving consciousness in intelligent agents. Recently this link between consciousness and emotions has been strengthened with the work by the neuroscientist Damasio [9] (see also R. Picard interview in this book pp.215). Thus, motivated by the role that emotions play in human cognition, McCauley et. al present two intelligent agents that were extended with emotional states. One of such agents, CMattie was designed to model a global workspace of consciousness and includes modules for perception, action selection, metacognition, associative memory, consciousness and emotions. In CMattie emotions are represented by emotion codelets (a small piece of code capable of performing some specific action under appropriate conditions) which influence current action selection, by influencing other codelets and behaviours of the agent. That is, the emotional states (with four emotions: anger, sadness, happiness and fear) will affect its cognitive processes influencing its reasoning and learning processes.

In Stevo Bozinovsky paper [5], emotions are used as elements in a learning based agent. The presented learning mechanism allows the definition of a "emotion learning agent" which combines both genetic elements and behavioral ones. Such an architecture was used to build a goal-keeper for a RoboCup [14] team element. The training of such a goal keeper showed that a good performance can be achieved using such an emotion-based architecture.

But, emotion synthesis can be used not only as a way to obtain more rational behaviour, but also to convey to the users some degree of believability. André et al. [1] present three systems (the Inhabited Market Place, the Puppet Project and the Presence System) where emotions and personality are used as a way of conveying some believability of the life-like characters. In the presented work, emotional behaviour is achieved through the use of two different emotion theories, in particular the OCC theory of emotions [20]. The systems presented and the results achieved are indeed a good example of how emotions and personality can be used effectively to convey believability of the characters.

4 The Expression of Emotions in Machines

In the past few years, several research teams have been enriching their human-computer interfaces with characters that exhibit facial and body expressions (see for example [15] [24] [2]). As argued in [4], creating an animated visual charac-

ter will be a common way to lend substance to our conversational computers. Such characters aim at making the communication between users and machines more effective, as the expressive power of the character can be combined with its utterances and actions. Further, such characters can also be enriched with emotional expressive power in order to both increase the believability as well as their interaction power.

P. Rizzo's paper (see [25]) discusses a set of hypotheses that are commonly assumed when giving synthetic characters the power of emotional expression. Those underlying assumptions are challenged and revisited in both a critical and constructive way. By discussing assumptions such as "the believability crucially depends on the agent's ability to show emotional behaviors" or "the ability to process affective states enables the agent to display emotional reactions; thus enhancing believability" Rizzo challenges the researchers on affect and believable agents to be more critical and to provide more evidences that believability, emotional expression and synthesis can still walk hands in hands.

Pat George and Malcolm McIlhagga (in [13]) discuss how emotional information can be communicated in a meaningful way to children using a virtual environment. They discuss the meaning that young children place on facial expressions when gathering information about emotion, and present an empirical study that shows how children discriminate among different emotional facial expressions in cartoon like characters. Further, they also examine how children use such emotional expressions to build creatively emotional characters acting within their stories.

At Microsoft, G. Ball and J. Breeze's research (see [4]) focuses on personality and its relation to external observable behaviors such as posture or gestures. Focusing on animated characters, and assuming that personality can also be conveyed by characteristic body motions, Ball and Breeze present a technique, based on Bayesian networks, by which such body expression comes as a result of the characters internal personality factors.

Poggi and Pelachaud (see [23]) direct their attention to the affect content of facial expressions and analyze the facial expression of emotion when contained in communicative acts, such as suggesting, warning approving, etc. They go further arguing that "the core meaning that is contained in the expression of an emotion is not only present in that emotion per se, but also in other kinds of communicative signals" (such as for example emphasis). So, they argue for the need to look at facial expression in animated characters through a compositional method in order to capture the richness of possible meanings that underly communicative facial expressions.

But affect does not only come in the form of facial or body expression. Emotional states bind the reasoning influencing the behaviour, the decisions, the speech, the choice of words, etc. In this book de Rosis and Grasso [10] present some elements of natural language generation that not only take into account the hearer's affective state, but mostly, is performed aiming at deliberately influencing his/her emotional states. They study the role that affect plays in the various stages of Natural Language Processing, and re-enunciate a set of prin-

ciples and assumptions which are then revised in order to create more "affective texts".

5 Final Remarks

This book combines different perspectives, theories and cases that show how to bring an affective dimension into the interaction between users and computer applications. It is my hope that the research presented in here can contribute greatly to an increased understanding of the role that affect plays in human/computer or computer-mediated interactions as well as for stimulating further research in this challenging new field.

References

1. Elisabeth André, Martin Klensen, Patrick Gebhard, Steve Allen, and Thomas Rist. Integrating models of personality and emotions into lifelike characters. In Ana Paiva, editor, *Affective Interactions: toward a new generation of computer interfaces (this volume)*. Springer, 2000. 5
2. Elisabeth André, T. Rist, and J. Müller. Guiding the user through dynamically generated hypermedia presentations with a life-like character. In *Proceedings of the 1998 International Conference on Intelligent User Interfaces*. ACM Press, 1998. 5
3. Luis Antunes and Helder Coelho. Redesigning the agents' decision machinery. In Ana Paiva, editor, *Affective Interactions: toward a new generation of computer interfaces (this volume)*. Springer, 2000. 5
4. Gene Ball and Jack Breese. Relating personality and behavior: Posture and gestures. In Ana Paiva, editor, *Affective Interactions: toward a new generation of computer interfaces (this volume)*. Springer, 2000. 5, 6
5. Stevo Bozinovski. Artificial emotion and emotion-learning: Emotion as value judgements. In Ana Paiva, editor, *Affective Interactions: toward a new generation of computer interfaces (this volume)*. Springer, 2000. 5
6. Cristino Castelfranchi. Affective appraisal vs cognitive evaluation in social emotions and interactions. In Ana Paiva, editor, *Affective Interactions: toward a new generation of computer interfaces (this volume)*. Springer, 2000. 4
7. Bridget Cooper, Paul Brna, and Alex Martins. Effective affective in intelligent systems- building on evidence of empathy in teaching and learning. In Ana Paiva, editor, *Affective Interactions: toward a new generation of computer interfaces (this volume)*. Springer, 2000. 2
8. António Damásio. *Descartes' Error*. Papermac, 1996. 1
9. António Damásio. *The Feeling of What Happens*. Heinemann, 2000. 5
10. Fiorella de Rosis and Floriana Grasso. Affective natural language generation. In Ana Paiva, editor, *Affective Interactions: toward a new generation of computer interfaces (this volume)*. Springer, 2000. 6
11. P. Ekman and Friesen W. *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, 1978. 1, 3
12. Nico Frijda. Varieties of affect: Emotions and episodes, moods, and sentiments. In Paul Ekman and Richard Davidson, editors, *The Nature of Emotion*. Oxford University Press, 1994. 1

13. Pat George and Malcolm McIllhagga. The communication of meaningful emotional information for children interacting with virtual actors. In Ana Paiva, editor, *Affective Interactions: toward a new generation of computer interfaces (this volume)*. Springer, 2000. 6
14. H. Kitano, M. Asada, Y. Kuniyoshi, I. Noda, and E. Osawa. Robocup: the robot world cup initiative. In *Proceedings of the first international conference on autonomous agents*. ACM Press, 1997. 5
15. J. Lester, S. Converse, S. Kahler, S. Barlow, B. Stone, and R. Bhoga. The persona effect: Affective impact of animated pedagogical agents. In Lewis Johnson and Barbara Hayes-Roth, editors, *CHI'97*. ACM Press, 1997. 5
16. Isabel Machado, Rui Prada, and Ana Paiva. Bringing drama into a virtual stage. In *Collaborative Virtual Environments 2000*. ACM Press, 2000. 4
17. Stacy Marsella, Lewis Johnson, and Catherine LaBore. Interactive pedagogical drama. In Carles Sierra, Maria Gini, and Jeffrey Rosenschein, editors, *Autonomous Agents 2000*. ACM Press, 2000. 2
18. Carlos Martinho, Isabel Machado, and Ana Paiva. A cognitive approach to affective user modelling. In Ana Paiva, editor, *Affective Interactions: toward a new generation of computer interfaces (this volume)*. Springer, 2000. 4
19. Lee McCauley, Stan Franklin, and Myles Bogner. An emotion-based "conscious software" agent architecture. In Ana Paiva, editor, *Affective Interactions: toward a new generation of computer interfaces (this volume)*. Springer, 2000. 5
20. Andrew Ortony, Gerald Clore, and Allan Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, 1988. 4, 5
21. Rosalind Picard. *Affective Computing*. The MIT Press, 1997. 1, 3
22. Rosalind Picard. Interview with rosalind picard: Author of "affective computing". In Ana Paiva, editor, *Affective Interactions: toward a new generation of computer interfaces (this volume)*. Springer, 2000. 2, 3
23. Isabella Poggi and Catherine Pelachaud. Emotional meaning and expression in animated faces. In Ana Paiva, editor, *Affective Interactions: toward a new generation of computer interfaces (this volume)*. Springer, 2000. 6
24. J. Rickel and L. Johnson. Integrating pedagogical capabilities in a virtual environment agent. In Lewis Johnson and Barbara Hayes-Roth, editors, *Autonomous Agents' 97*. ACM Press, 1997. 5
25. Paola Rizzo. Why should agents be emotional for entertaining users? a critical analysis. In Ana Paiva, editor, *Affective Interactions: toward a new generation of computer interfaces (this volume)*. Springer, 2000. 6
26. Dorothy G. Singer and Jerome L. Singer. *The House of Make Believe*. Harvard University Press, 1990. 2
27. Alan Sroufe. Socioemotional development. In Joy D. Osofsky, editor, *Handbook of Infant Development*. New York: Wiley, 1979. 2
28. Thomas Wehrle and Susanne Kaiser. Emotion and facial expression. In Ana Paiva, editor, *Affective Interactions: toward a new generation of computer interfaces (this volume)*. Springer, 2000. 3, 4
29. Gillian M. Wilson and Angela Sasse. Listen to your heart rate: Counting the cost of media quality. In Ana Paiva, editor, *Affective Interactions: toward a new generation of computer interfaces (this volume)*. Springer, 2000. 2, 4

Listen to Your Heart Rate: Counting the Cost of Media Quality

Gillian M. Wilson and M. Angela Sasse

Department of Computer Science, University College London
Gower Street, London, WC1E 6BT
g.wilson@cs.ucl.ac.uk
a.sasse@cs.ucl.ac.uk

Abstract. Videoconferencing over the Internet is developing into a valuable communicative tool, therefore assessment methods allowing users' requirements to be determined and accounted for in the design of applications are vital. Subjective rating scales are mainly used to assess whether multimedia quality is sufficient for a particular task, however relying on this method alone has drawbacks. Therefore, we are investigating the use of objective methods to assess the user cost of different levels of multimedia quality: physiological indicators of stress are being measured. Two aims of this research are 1) to determine the optimum and minimum levels of quality which different users require for performing different tasks without significant user cost and 2) to produce a utility function which will allow the application to receive physiological feedback on the user.

1 Introduction

The number of networked multimedia applications, like multimedia conferencing (MMC), increases constantly. MMC facilitates communication between two or more users through audio, video and shared workspace tools in real time. It is becoming more widespread due to falling costs of the hardware required and improvements in networks. MMC is viewed as valuable in a large number of areas, such as distance learning and distributed project collaboration.

It is possible to send and receive audio and video of a high quality, yet this potentially gives rise to an increase in its financial cost for the user, who will not want to pay more than is needed for this medium of communication. Therefore, the levels of media quality that allow users to complete their task effectively and comfortably are essential information for network providers and application designers.

At present, the most common way of assessing multimedia quality utilises subjective assessment. However, relying on this alone has drawbacks. Therefore, this paper presents a new method to assess the quality of networked applications: physiological indicators of stress to media quality are being measured as an indicator of user cost. These measurements are taken as part of a traditional HCI (Human

Computer Interaction) assessment framework that also includes measures of task performance and user satisfaction. Using this approach will give a better indication of quality thresholds required by users. In addition there is the possibility of feeding real-time physiological data into the application, which could then modify itself if the user was under stress. Thus, providing an ‘affective’ interaction between the computer and user.

2 Assessment of Media Quality

The ITU (International Telecommunications Union) subjective rating scales are most commonly used to assess quality in this area. These involve a short section of material being played, after which a 5-point quality/impairment rating scale is administered and a Mean Opinion Score (MOS) calculated. However, recent research has raised concerns about their effectiveness in evaluating multimedia speech [24], and video. The main problems are:

- The scales are one-dimensional, thus they treat quality as being uni-dimensional. This approach is questionable as there are many factors which are recognised to contribute to users perception of audio [12] and video [10] quality.
- They were designed to rate toll quality audio and high quality video: MMC audio and video are subject to unique impairments.
- The scales are mostly concerned with determining if a viewer/ listener can detect a particular degradation in quality, whereas with MMC it is more important to determine if the quality is *good enough* for the task.
- The short duration of the scales means that there is not the opportunity for the viewer/listener to experience all the degradations that impact upon MMC. Subsequently, a dynamic rating scale for video is now recommended by the ITU (ITU- BT 500-8)[11] in order to account for changes in network conditions.
- The vocabulary on the scales (Excellent, Good, Fair, Poor, Bad) is unrepresentative of MMC quality and the scales are not interval in many languages, therefore scores obtained can be misleading.

In order to address these problems, an unlabelled rating scale was devised [23] and studies showed that users were consistent in their quality ratings using the scale. A dynamic software version of this scale was subsequently developed, QUASS (QQuality ASsessment Slider), which facilitates the continuous rating of the quality of a multimedia conference [3]. The drawback with this method is that continuous rating can result in task interference.

There are also fundamental problems associated with subjective assessment methods, which center on the fact that they are cognitively mediated. For example, it was discovered that users accepted significantly lower levels of media quality when financial cost was attached: the accepted quality levels were below the threshold previously established as necessary for the task [2]. Another example of cognitive mediation was given by Wilson & Descamps [26], who showed that the level of task difficulty can affect the rating given to video quality: the same video quality received a lower quality rating when the task was difficult. Therefore, it can be concluded that

users may not always be able to accurately determine/judge the quality they need to complete a particular task when contextual variables are operating.

Finally, Knoche et al [13] argue that subjective methods are fundamentally flawed as it is not possible for people to register what they do not consciously perceive, thus they recommend that measures of task performance are used to determine how good or bad quality is. We believe that whilst both subjective assessments of user satisfaction and task performance are extremely important in the evaluation of multimedia quality, to rely solely on one method would be flawed. Therefore, we have adopted the traditional HCI approach of measuring task performance, user satisfaction and user cost.

3 Objective Measures of Users' Responses to Quality

We set out to investigate the use of objective methods to identify when inadequate media quality is causing stress: this type of measurement is known in HCI as *user cost* and has largely been neglected in this area. Thus, here user cost is defined as stress. Stress is a disruption in the homeostasis of an individual due to disagreeable stimuli.

If a user is performing a task under low quality audio and video, he/she must expend extra effort at the perceptual level in order to decode the information. If the user has difficulty in decoding this information, then they should experience discomfort or stress, even if they can still complete their task. It is likely that any effect on task performance and user satisfaction would only arise in the long-term, as people may not realise that they are under stress in short lab-based experiments when engaged in a task (see section 6.2).

Stress is being examined for the following reasons. Firstly, it is an instant, unconscious and automatic mechanism that is not under the influence of cognitive mediation. Secondly, its measurement physiologically does not interfere with task completion. Thirdly, stress is widespread in the workplace, thus any attempt to reduce its amount is beneficial. Finally, stress can have a very negative impact on people, especially those who suffer from chronic stress. Such prolonged stress can give rise to illness by having an impact on the immune system.

3.1 The Measurement of Stress

To measure stress, the following signals have been adopted: Blood Volume Pulse (BVP), Heart Rate (HR) and Galvanic Skin Resistance (GSR). These signals were chosen as they are unobtrusive, are good indicators of stress and are easy to measure with specialised equipment (see section 3.3).

The nervous system of humans is separated into the central nervous system (CNS) and the peripheral nervous system (PNS). The PNS comprises the somatic nervous system (SNS) and the autonomic nervous system (ANS). These both interact with the CNS. The function of the SNS is to carry the messages to and from the muscles controlling the skeleton whereas the ANS carries messages to and from the body's internal organs.

It is the ANS that is concerned with the regulation of internal structures, which are smooth muscle i.e. in blood vessels and skin, heart muscle and glands. Autonomic means ‘self-regulating’, thus describing the fact that most of its structures are not under conscious control. The ANS is divided into the sympathetic and the parasympathetic divisions.

The sympathetic division activates the body’s energetic responses. When faced with a stressful situation the CNS will make the conscious decision whereas the ANS immediately mobilises itself without the need for conscious instruction. This is often referred to as the ‘fight or flight’ response [5]. The sympathetic division prepares the body for action by e.g. speeding up the heart rate, dilating the walls of the blood vessels to speed up blood flow to the limbs and releasing glucose into the bloodstream for energy. These actions are reinforced by the endocrine system because sympathetic nerves also directly stimulate the adrenal glands to release adrenaline.

When the stressful situation has passed, the parasympathetic division takes over to restore the body’s balance. The purpose of this is to conserve the body’s resources. Therefore it prompts salivation and gastric movements and encourages blood flow to the gastrointestinal system and storage of glucose by the liver.

3.2 What Happens to These Responses under Stress?

Heart rate is viewed as a valuable indicator of overall activity level, with a high heart rate being associated with an anxious state and vice versa [8]. Seyle [20] has linked GSR to stress and ANS arousal. GSR is also known to be the fastest and most robust measure of stress [4], with an increase in GSR being associated with stress. BVP is an indicator of blood flow: the BVP waveform exhibits the characteristic periodicity of the heart beating: each beat of the heart forces blood through the vessels. The overall envelope of the waveform pinches when a person is startled, fearful or anxious, thus a decrease in BVP amplitude is indicative of a person under stress and vice versa.

Under stress, HR rises in order to increase blood flow to the working muscles, thus preparing the body for the ‘fight or flight’ response. GSR increases under stress: the precise reason this happens is not known. One theory is that it occurs in order to toughen the skin, thus protecting it against mechanical injury [25] as it has been observed that skin is difficult to cut under profuse sweating [6]. A second theory is that GSR increases to cool the body in preparation for the projected activity of ‘fight or flight’. BVP decreases under stress. The function of this is to divert blood to the working muscles in order to prepare them for action. This means that blood flow is reduced to the extremities like a finger or a toe.

3.3 How Are These Responses Being Measured?

The equipment being used to measure the physiological responses is called the ProComp and is manufactured by Thought Technology Ltd. [21]. It can monitor up to eight channels simultaneously, thus the option is there for this research to build on the measure set currently being used. Additionally, the ProComp can be used for biofeedback (see section 7.2). In order to measure GSR, two silver-chloride electrodes

are placed on adjacent fingers and an imperceptible small voltage is applied. The skin's capacity to conduct the current is measured. The technique of photoplethysmography is used to measure HR and BVP. This involves a sensor being attached to a finger or a toe. The sensor applies a light source and the light reflected by the skin is measured. At each contraction of the heart, blood is forced through the peripheral vessels, which produces an engorgement of the vessel under the light source. Thus, the volume and rate at which blood is pumped through the body are detected.

4 Research Questions

There are many issues this research is tackling. Firstly, what aspects of objective quality delivered can be stressful? Factors affecting audio and video quality due to the network, like packet loss, will be investigated along with variables dependent on the end system, such as the color depth of video.

Secondly, how can stress and emotions, such as excitement about the situation or task, be separated in an experiment? This is a problem, as the physiological patterns accompanying emotions have not been mapped [4], yet recent research at MIT showed that the physiological responses of eight emotions can be distinguished between with 80% accuracy [22], which is an encouraging result. However, in the experiments used in this research, we need to ensure that any stress observed is due to the quality, as opposed to any other factors. To address this we use the following methods:

- Baseline responses are measured for fifteen minutes prior to any experimentation so that a 'control' set of data is available with which to compare responses under quality levels, and to allow the sensors and measurements to settle down.
- The first five minutes of physiological responses in experiments are discarded in order to ensure that the change from baseline measurements being taken to the experiment commencing is not affecting results.
- The environment of the experiment is held constant to ensure that events, such as the phone ringing, are not affecting users.
- The tasks used are carefully designed to ensure that they are not overly stressful, yet remain engaging.

Thirdly, what physiological values indicate that a user is under stress? The responses that are significantly different from the baseline measurements and in the direction of stress are taken to indicate that the user is under stress. This accounts for the fact that everyone has different physiological responses and baseline values. To cross-validate results, we utilise subjective assessment to determine if the participants say they felt under stress during the experiment.

Fourthly, can stress ever be viewed as beneficial to performance? It has been found that stress, in some situations, can actually enhance performance e.g. by improving memory due to the facilitatory effects of adrenaline [14]. Therefore, it could be argued that stress might be beneficial in a working environment. However, a technology with which people come into contact every day and one which causes

stress is negative: people cannot function effectively under stress for a prolonged period and it can actually be harmful to the body (section 3) and performance e.g. prolonged stress can devastate explicit memory [14].

Fifthly, how is stress related to emotions? This question is one which is under contentious debate at present, yet for our research this is not of paramount importance. Whilst it is a very significant and interesting issue, we are more concerned with the physiological detection of stress and the implications this has for media quality requirements: determining if it is an emotion is not critical for this purpose.

Finally, is this a reliable way to measure user cost? Only further research will provide the answer.

5 Affective Computing

Affective computing is a field currently attracting interest. The term refers to computing that is “related to, arises from or deliberately influences emotions” [17]. The main focus is to create computer systems that can identify, comprehend and even ‘have’ human emotions as it is now accepted that emotions are critical in activities such as decision making, perception and learning. Research into building computers with these attributes is being conducted at the Media Laboratory of MIT (Massachusetts Institute of Technology).

The main area of relevance of Affective Computing to this research is in the area of sensing human emotions. Researchers in the Media Lab, i.e. [18], are attempting to build wearable computers using discrete lightweight sensors that can be worn for long periods of time without discomfort. Such a development would be extremely beneficial to this research, as it would allow the long-term monitoring of users physiological responses in field trials in an inconspicuous manner. Additionally, it would allow the computer to obtain continuous physiological information on the state of the user upon which it could then act (see section 7). Importantly, the research being conducted at MIT is promoting the use of physiological measurements as indicators of stress.

6 Experimental Data

6.1 Pilot Trial

As a starting point to the research, a small pilot trial was conducted to determine if it would be possible to detect poor audio quality physiologically. Participants listened to audio segments that had varying loss rates. GSR, BVP and HR were measured throughout the trial. We hypothesised that HR and GSR would be highest, whereas BVP would be lowest in the condition with the worst loss rates, thus indicating stress.

It was not valid to perform statistical analyses, as there were only six participants. However, the mean responses showed that GSR and HR responded as predicted, whereas BVP did not. This result illustrated that it was possible to obtain

physiological responses to poor audio quality from at least two of the signals being used. However, would statistically significant results arise in a full experiment and one that examined video? It is accepted that in multimedia conferences audio plays the more important role [19].

6.2 Full Experiment

Anderson et al [1] used subjective assessments and measures of task performance to determine if users could determine the difference between 12 and 25 video frames per second (fps) when they were undertaking an engaging task. Results showed that users did not notice the change in frame rate and it had no effect on task performance. However, O'Malley et al [16] found that the same difference is noticed when the data are short video clips played isolation. If users do not notice the difference in frame rate when engaged in a task, does this imply that it has no effect on them? Such a finding would have positive implications for the saving of bandwidth.

To investigate this issue, an experiment looking at two frame rates, 5 and 25 fps (increasing the quality difference used in [1]) has taken place. Twenty-four participants watched two recorded interviews conducted using IP (Internet Protocol) videoconferencing tools on a high-quality computer screen. The interviews were between a university admissions tutor and school pupils, who played themselves in scripted interviews. The script was designed with the assistance of an admissions tutor and aimed to emulate common exchanges of information in such interviews, whilst keeping the content minimally stressful.

The interviews lasted fifteen minutes each. Participants saw two interviews at 5-25-5fps or 25-5-25fps, where each frame rate was held for five minutes, and were asked to make a judgement on the suitability of the candidates. The frame rate changed twice to ensure that no expectancy effect would arise. Audio quality was good and did not change.

Participants rated the audio/video quality using the QUASS tool. After the interviews a questionnaire was administered. This addressed how participants felt during the experiment and their opinions on the quality. Physiological measurements were taken throughout the experiment. We posited the following hypotheses:

1. There will be different physiological responses to the two frame rates: 5fps will cause more stress.
2. Participants will not register the frame rate change subjectively.

T-tests were performed on the physiological data: 75% of participants mean GSR, HR and BVP showed a statistically significant increase in stress at 5 fps (GSR, $p=0.002$; HR, $p=0.003$; BVP, $p=0.04$). Figure 1 shows the mean heart rate of one participant over three frame rates. There was no statistically significant correlation between subjective (QUASS) and physiological responses. In addition, questionnaire analyses showed that only 16% of participants noticed that the frame rate had changed. Thus, both hypotheses were supported.

These results are important as they indicate that when people are engaged in a task, they say they do not notice the difference between two different frame rates during or after the task, however the difference is registered physiologically. This

implies that high frame rates should be provided in order to reduce user cost. In addition, acceptable quality levels required for a task and those which result in unacceptable user cost should not be determined by subjective assessment alone: we recommend that objective data is also collected.

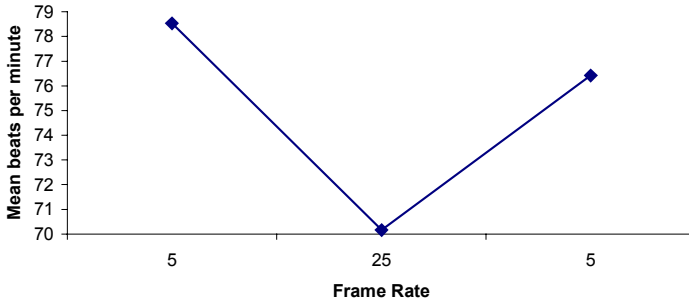


Fig. 1. Graph showing the mean heart rate of one participant over three frame rates

7 Conclusions and Future Work

Traditional HCI approaches to evaluation recommend that task performance, user satisfaction and user cost should be considered: we feel that in the evaluation of media quality this is particularly appropriate. This paper has shown that 1) media quality does impact physiological responses 2) subjective assessments used in isolation do not give a true indication of the impact of quality upon the user 3) user cost should be included in multimedia quality evaluation and in wider areas such as product assessment. The research detailed in this paper will produce three substantive contributions:

- The optimum and minimum levels of some parameters of multimedia quality required for users performing tasks without significant user cost will be determined. This will aid network providers in determining resource allocation and will ultimately benefit the end-user, as their needs will be more clearly specified.
- This research feeds directly into the ETNA Project (Evaluation Taxonomy for Networked Multimedia Applications) [7], which aims to produce a taxonomy of real time multimedia tasks and applications. The maximum and minimum audio/video quality thresholds for a number of these tasks will also be determined. This will assist network providers and application designers, as guidelines of users' quality requirements for specific tasks will be provided.
- A utility curve, a mechanism by which the network state can be directly related to the end-user, will be created. This will allow the application to receive feedback on the user. Wearable computers would be useful here to feed physiological information into an application like MMC. If the application detected that the user was under stress, it would adjust the variable of the multimedia conference

causing stress in order to increase user satisfaction. If the network could not supply the optimum level of quality, then the utility curve would be consulted to provide the next best quality level and so on. If no suitable level of quality could be provided due to e.g. network congestion, the application would then inform the user and end the session. It is likely that the user would be placed under too much stress by such poor quality levels and this would ultimately affect task performance and user satisfaction. Thus, the application would be taking into account the physiological state of the user and providing an 'affective interaction'.

Finally, a methodological contribution will be made: guidelines on the most appropriate physiological measurements to indicate a particular impairment in quality will be produced. This will pave the way for much needed further research in this area.

7.1 Moving the Research into the Field

Customer service is cited as the key to the successful company of the future. The call centre industry is currently experiencing a boom, yet customers' perceptions of services by telephone are declining. Thus, it is vital that fast accurate information with prompt solutions is provided to the customer and that a good customer relationship is formed. However, the fact that there is a 30% annual turnover in call centre staff indicates that the job can be stressful and repetitive. How can such a job be improved for the employees so that the effects ultimately pass down to the customers?

Recent research at Bournemouth University looked at how elements that motivate customer services employees can be incorporated into the computer desktop. Subsequently, the MUI (Motivational User Interface) was developed between Bournemouth University and British Telecom (BT) Bournemouth '150' call centre [15]. The MUI allows call centre operators to:

- personalize their workspace
- lets them know how busy they are
- gives them the ability to effectively manage customer information in the format in which the customer sees it
- incorporate elements of moral support and teamwork
- unleash their frustration

We are currently in discussions with BT about the possibility of utilising physiological measurements, which could include muscular-skeletal stress, in two areas. Firstly, as a method to evaluate the interface and secondly to feedback physiological data into the application.

Physiological stress would be detected and measures of user satisfaction and task performance would be administered in order to determine how experienced users of the MUI worked using the new interface. These results would then be compared to those of operators using the traditional interface, in order to obtain an indication of which interface is more effective and induces the least cost for the user.

There is also the possibility of making the MUI adaptive to the physiological measurements, in effect sympathising with the user. For example, if a user was eliciting stress responses the interface could e.g. change the colour of the screen to induce calm, or bring up a web browser to allow them to have a short break from the job.

The interest from the MUI team in this research illustrates the potential for moving these measurements out of MMC assessment and into user-centric technology evaluation and assessment as a whole.

7.2 More General Applications

Stress is extremely common in today's workplace. Thus, the technology that people use daily should be designed to cause them minimal stress. Some ideas of taking this research further are that it could give the user more self-awareness about their stress levels, thus encouraging them to 'self-manage' stress. This has links with the discipline of biofeedback, which is a technique for monitoring physiological activity and converting it into an auditory or visual message [9]. It has been shown that e.g. tension headaches and heart rate can be controlled by giving feedback without any external reward as it seems like the feedback and perceived control are intrinsically rewarding. Biofeedback is often used to decrease stress.

It may also be beneficial for bosses to observe their employees stress levels, in order to make them more aware of the pressures placed upon employees. Additionally, the detection of stress could be beneficial for illnesses like RSI (Repetitive Strain Injury). For example people who do a lot of a specific activity, like typing, could wear muscle sensors to alert them to take a break and to rest before any lasting damage is done. This would be extremely useful when it is considered that if a user is involved in an engaging task, they may not notice the stress they are under, as was illustrated in our full video trial (see section 6.2).

In conclusion, the ability to detect stress unconsciously has wide ranging implications from areas like product assessment to providing 'emotionally sympathetic' user interfaces.

Acknowledgments

We gratefully acknowledge the contribution of Anna Bouch from UCL Computer Science. Gillian Wilson is funded through an EPSRC CASE studentship with BT Labs.

References

1. Anderson, A.H., Smallwood, L., MacDonald, R., Mullin, J. and Fleming, A. (in press 1999/2000) Video data and video links in mediated communication: What do users value? *To appear in International Journal of Human Computer Studies*.

2. Bouch, A. and Sasse, M.A. (1999) Network quality of service: What do users need? 4th International Distributed Conference. Madrid, 22nd-23rd September 1999.
3. Bouch, A., Watson, A. and Sasse, M.A. (1998) QUASS – A tool for measuring the subjective quality of real-time multimedia audio and video. *Proceedings of HCI '98*, 1-4 September 1998, Sheffield, UK.
4. Cacioppo, J.T. and Louis, G.T. (1990) Inferring psychological significance from physiological signals. *American Psychologist*, 45(1): 16-28.
5. Cannon, W.B. (1932) *The Wisdom of the Body*. (Reprinted 1963.) New York: WW Norton
6. Edelberg, R and Wright, D.J. Two GSR effector organs and their stimulus specificity. Paper read at the Society for Psychophysiological Research, Denver, 1962.
7. ETNA Project. <http://www-mice.cs.ucl.ac.uk/multimedia/projects/etna/>
8. Frijda, N.H. (1986) The Emotions, chapter Physiology of Emotion, p124-175. *Studies in Emotion and Social Interaction*. Cambridge University Press, Cambridge, 1986
9. Gatchel, R. (1997) Biofeedback. In A. Baum, S. Newman, J Weinman, R. West and C McManus. *Cambridge Handbook of Psychology, Health and Medicine*. Cambridge: Cambridge University Press.
10. Gilli Manzanaro, J., Janez Escalada, L., Hernandez Lioreda, M. and Szymanski, M. (1991) Subjective image quality assessment and prediction in digital videocommunications. COST 212 HUFIS Report, 1991.
11. ITU-R BT.500-8 Methodology for the subjective assessment of the quality of television pictures: <http://www.itu.int/publications/itu-t/itu-rec.htm>
12. Kitawaki, N. and Nagabuchi, H. (1998) Quality assessment of speech coding and speech synthesis systems. *IEEE Communications Magazine*, October, 36-44, 1988.
13. Knoche, H., De Meer, H.G. and Kirsh, D. (1999) Utility curves: mean opinion scores considered biased. *Proceedings of 7th International Workshop on Quality of Service*, 12-14, 1999
14. LeDoux, J. *The Emotional Brain*. New York: Simon and Schuster, 1996
15. Millard, N., Coe, T., Gardner, M., Gower, A., Hole, L. and Crowle, S. (1999) The future of customer contact. *British Telecom Technology Journal*. <http://www.bt.co.uk/btj/vol18no1/today.htm>
16. O'Malley, C., Anderson, A.H., Mullin, J., Fleming, A., Smallwood, L. and MacDonald, R. Factors Affecting Perceived Quality of Digitised Video: Tradeoffs between Frame Rate, Resolution and Encoding Format. Submitted to *Applied Cognitive Psychology*, forthcoming.
17. Picard, R.W. *Affective Computing*. M.I.T. Press, Cambridge, MA, 1997.
18. Picard, R.W. and Healey, J. (1997) Affective wearables. In *Proceedings of the First International Symposium on Wearable Computers*, Cambridge, MA, Oct. 1997
19. Sasse, M.A., Biltung, U., Schulz, C-D. and Turletti, T. (1994a) Remote seminars through multimedia conferencing: experiences from the MICE project. *Proceedings of INET'94/JENC5*.
20. Seyle, H. *The Stress of Life*. McGraw-Hill, 1956

21. Thought Technology Ltd. <http://www.thoughttechnology.com/>
22. Vyzas, E. and Picard, R.W. (1999) Offline and online recognition of emotion expression from physiological data. Workshop on Emotion-Based Agent Architectures, Third International Conference on Autonomous Agents, Seattle, WA, 1999.
23. Watson, A and Sasse, M.A. (1997) Multimedia conferencing via multicast: determining the quality of service required by the end user. Proceedings of AVSPN '97 – International Workshop on Audio-visual Services over Packet Networks, 189-194, 1997.
24. Watson, A and Sasse, M.A. (1998) Measuring perceived quality of speech and video in multimedia conferencing applications. *Proceedings of ACM Multimedia '98*, Bristol, UK, September 1998. ACM New York, pp.55-60.
25. Wilcott, R.C. (1967) Arousal Sweating and Electrodermal Phenomena. *Psychological Bulletin*, 67, 58-72.
26. Wilson, F. and Descamps, P.T. (1996) Should we accept anything less than TV quality: Visual communication. International Broadcasting Convention, 12th-16th September 1996.

Effective Affective in Intelligent Systems – Building on Evidence of Empathy in Teaching and Learning

Bridget Cooper, Paul Brna, and Alex Martins

Computer Based Learning Unit, Leeds University
Leeds LS2 9JT, England

{bridget,paul,alex}@cbl.leeds.ac.uk
<http://www.cbl.leeds.ac.uk/~bridget/>

Abstract. This paper considers how research into empathy in teaching and learning can inform the research into intelligent systems and intelligent agents embedded in educational applications. It also relates this research to some analysis of classroom practice completed as part of the EU funded NIMIS project. The project is developing three applications, one of which aims to support writing development with young children aged 5-6 years based on a cartoon format. The NIMIS classroom as a whole is designed to enhance and augment existing classroom practices and to foster collaboration by non-intrusive hardware and intuitive hardware and software interfaces. To this end it seeks to enhance both human and electronic communication in the classroom. Empathy is central to ensuring the quality of human communication and personal development. This paper suggests that intelligent systems that can consider more carefully the processes and feelings involved in human interactions in teaching and learning, may promote higher quality support for students in classrooms.

1 Introduction

This paper considers how research into empathy in teaching and learning can inform the research into intelligent systems and educational applications. It also relates this research to the NIMIS project, one of the current i3 projects in Experimental School Environments. The main objective of NIMIS is to develop a classroom for early learners, featuring a synergy between social, educational and technological factors. This will be achieved through a range of classroom activities oriented towards learning to comprehend and manipulate a range of media (text, pictures, animation, gesture, and sound). These activities take advantage of the possibilities created through a unique combination of hardware and highly interactive software. The NIMIS team believes that education of young children can be improved by a new kind of classroom, one which respects that children need to learn in the company of other children and their teacher, while taking advantage of adaptive and highly interactive learning environments. Young pupils will experience enhanced classroom interaction through using networked

computers together with a large interactive display (electronic board) in a framework called the Computer integrated Classroom (CiC). The project is developing three applications, which aim to support literacy development and collaborative skills with young children aged 5-8 years. The NIMIS classroom as a whole is designed to enhance existing classroom practices and to foster collaboration by non-intrusive hardware and intuitive hardware and software interfaces. To this end it seeks to support both human and electronic communication in the classroom. Empathy is central to ensuring the quality of human communication and personal development. This paper suggests that intelligent systems that consider more carefully the processes and feelings involved in human interactions in teaching and learning, may provide higher quality support for students in classrooms.

2 Educational Framework

The increasingly technicist approach to education in the eighties and nineties has produced an inevitable response from the educationalists who prefer to visualise a whole child, a small human being, in the educational process, rather than a pupil to whom information is transmitted. The revival of interest in the more human aspects of institutions, in both the educational and business world, through the development of emotional intelligence [9] is also testimony to this. Despite government pressure for tightly controlled learning regimes, tick lists of teaching competences, clearly defined standards and a strong emphasis on subject knowledge in teacher training, especially in the United Kingdom at present, research tends to suggest that this is not necessarily the route to the most effective teaching and learning. Ironically, much research evidence suggests that higher standards in education are achieved by the power of positive human relationships engendered between individuals in communities with shared values [20,1,22]. This is not to say that subject knowledge is not important. On the contrary, it is vital. However, unless it is harnessed to personal and communal relationships it can be perceived by children as detached, unattractive and irrelevant. Unless knowledge holds personal meaning for the child and can be internalised, it may well remain inaccessible and non-transferable [12].

Meaning is constructed through the continuous dialogue which takes place in the teaching and learning process in all its various forms [27]. According to Heathcote, learning is imbued with significance when one human being talks to and listens to another [14]. Children come to classrooms with prior learning and understanding, which has to be assessed, built on and incorporated into their new frameworks of understanding [2, pg 50]. Central to learning therefore, is not just the child's ability to understand the teacher but the teacher's ability to understand the child and the diverse range and nature of understanding in different children. This requires a considerable degree of empathy. This 'sensitive understanding' (a phrase used by Carl Rogers in 1975 [26]) takes into account emotional investment in learning and personal circumstances and peer relationships as well as existing schemes of knowledge. Such understanding is dependent

on the teacher's ability to empathise with her students in all their different circumstances. Only by developing a rich understanding of her pupils and offering diverse teaching methodology, at different attainment levels and with different personal interaction, can the teacher properly scaffold learning and meet varying learning needs. By targeting need effectively and facilitating success for different pupils at different levels, the teacher builds the crucial self-esteem, which enables children to take new risks in their understanding, allows them to modify their existing beliefs and creates the atmosphere in which rapid learning can take place.

3 Empathy in Intelligent Systems

Though an artificial tutor coupled with a believable agent cannot really empathise with or understand the students to whom it responds, it can demonstrate empathic characteristics which improve the learning climate and help to meet the individual learning needs of students. Research in artificial intelligence (AI) and computer mediated communication (CMC) has already identified a range of empathic characteristics. These characteristics include: positive affirmation and understanding [29]; motivating aspects of facial expression, voice tone and body language [23]; and the creation of a persona with empathic responses in combination with knowledge-based learning environments [17]. These varying elements help to identify and meet learning needs in the right way at the right time. Recent research into children's responses to avatars showed the particular importance of the listening response and the significance of 'envelope behaviours' [3], which are also shown to be key empathic characteristics in the recent research described later in this paper.

However at times, the traditional, real world models of teaching and learning, on which some artificial environments are modelled are less empathic than they might be, based as they often are on a more traditional classroom teacher approach. The commonly experienced 'teacher' model, which all developers have experienced themselves at school and which must influence their thinking, is not an ideal model for high quality learning. In a sense what we need to model our agents on, is not so much the 'common' experience of teaching, as a best practice model, which may be significantly different. Even if we had an accurate concept of and personal experience of an exceptional classroom teacher (about which there is limited current research), the well-known and frequently experienced concept of a teacher is nearly always less than ideal. This is because teaching and learning takes place, for the most part, in large classes, under relatively formal and often pressured conditions, with teachers spending a large part of their time managing and controlling, rather than being engaged in sensitive teaching and learning. The 'factory' system of education commonly experienced both by children and older students world-wide, does not create prime conditions for teaching and learning and teachers themselves understand only too well, the inadequacy of the conditions in which they normally work.

Though teachers try to make the best of the constraints in schools, there is an essential conflict between crowd control and learning. This conflict turns good classroom teaching and all the complexity that it involves, into a highly skilled art, but one which frequently confuses classroom management with teaching and learning. Though we can observe good teachers at work in classrooms, perhaps we also need to produce a more accurate model of a good tutor in one-to-one or small group situations. Intelligent tutors have the distinct advantage of working one-to-one and we need to understand what high quality one to one teaching can be like, rather than the more frequently seen one to thirty technique, which is inevitably very different. If we remove the ‘crowd control’ elements, perhaps we can come closer to modelling a more effective personal tutor. Equally, if we can understand some of the constraints and pressures on the teacher, created by managing large groups, we may be better able to capitalise on the strengths of the artificial support and provide more appropriate support in the classroom as a whole.

A few examples of the kinds of behaviour, embedded in intelligent systems, which tend to emerge from the ‘one to thirty’ scenario are such things as:

- making students wait for a response .This is inevitable in a classroom but can be perverse in a one to one situation.
- the use of facial expression and body language suggesting impatience and frustration with the student. This is typically found in a classroom where a teacher is under severe time constraints and required to manage large groups of learners, with varying levels of attainment, through a prescribed and, at times, inflexible curriculum.
- a strongly instructive approach to learning. This is a process often forced through lack of time when dealing with large numbers.
- a tendency to over control the entire learning process. This is a feature brought about by the need for the traditional teacher to run a whole class and manage multiple students and constraints in a complex environment.

The complex task of running a whole classroom tends to shift the balance in the classroom away from learning and towards control. In a sense, an empathic relationship demands a more but not entirely equal role between student and teacher, one where the teacher allows freedom where it is appropriate but also gives support and direction when it is needed. One response by researchers in artificial intelligence to these inherently inadequate but commonly experienced tutor models, seems to have been to turn to the peer model instead, under the auspices of collaborative learning [25,13]. Though this is viable and understandable in terms of offering a more equal relationship, which initially may appear more empathic and is important for learning, the peer model can also present problems in terms of meeting individual needs, if it is used as the key support. Whereas a peer might be helpful to a fellow student, they are not in the same position of responsibility to another student in the way that a teacher is. Given a choice between meeting their own and another’s needs, they are quite likely to choose their own and are unlikely to have the time to devote to another student

with considerable problems. Peers frequently compare themselves with and compete with one another and the education system and society at large positively encourages this. This makes peer co-operation and support difficult to ensure. Neither do peers generally have the same experience and expertise as the teacher in order to give the most appropriate scaffolding. Peer ‘support’ can at times be unhelpful if there is no appropriate teacher intervention.

Complex peer relationships can be taken into account by an expert teacher working with a group and they can then make appropriate decisions and responses which recognise and compensate for differences and problems. For example, dominating students may be encouraged to listen more and shy students encouraged to express their views. However with no structure or intervention or variety of approach, group activities can rapidly deteriorate into a battle for supremacy, which some students are ill-equipped to handle. If the values supported by artificial intelligence are to be inclusive then over-reliance on peer tutoring, may produce an inadequate model.

The problems raised by peer discussion led AI to try various computer interventions in the role of facilitator with sentence openers and suggestions for initiating dialogue [29,25] in conjunction with considerable research into dialogue analysis. Dialogue support is helpful in some circumstances and encourages student involvement but it is also limited because it often avoids the adoption of a directive or expert interventionist role, which is also necessary for learning. ‘Contingent teaching’ described by David Wood [28], requires careful monitoring and intervention by the tutor. Particularly with young children or new students or students who are short of confidence, a teacher needs to know when to give very concrete and specific support to promote development or perhaps to prevent a student being distracted from completing a larger task. For example, for a child who is a confident reader, it might be appropriate to ask them to sound out a longer word, in the middle of a story, as it would not interrupt their flow. However, a less confident child or more novice reader might find this disruptive, so the teacher might quickly read the word for the child to maintain the flow. Being able to give support can be as important as being able to stand back; the important element is knowing when to do both. These decisions are very finely tuned and dependent on the teacher having a rich understanding of each student both emotionally and academically. One student for example, could take correction more easily than another. For some students, criticism could be devastating and inhibit further learning. Student responses can vary between and within sessions making rules and guidelines for positive dialogues difficult to extricate. An observant tutor will notice the effect when they or a peer make an inappropriate response and can subsequently compensate for it.

Increasingly research in AI is examining the teaching and learning process in a more complex way, (sometimes in a more complex way than traditional educators). It has considered switching roles in response to the student, taking into account both the affective and cognitive elements in learning [10] and the combination of peer and tutor roles [13]. This complex approach may be perhaps, the most fruitful area to explore. In fact the recent research into empathy out-

lined below suggests that teachers switch between different roles very quickly in their relationships with classes and individual pupils. Teachers continually make spontaneous decisions in the classroom and adopt different styles, voice tone and approaches in order to meet individual and group needs. They may play a directive role in one instance, switching to an empowering, choice giving role the next. They may oscillate between a fun-loving companion, a critical friend or an expert advisor on domain specific knowledge. They may be the person to whom a child feels able to relate a joke or personal story or they may take time to reveal their own feelings, passions or interests and this too creates an ambiance of trust, which is vital to learning. It is this multiplicity of roles in combination which properly ensures the meeting of diverse needs and which is also indicative of advanced empathy. If our systems are going to support learning then they may need to be multifaceted and responsive to a wide range of different needs. Some of these issues have been identified by AI research as being directly related to empathy in the teacher/pupil relationship [19]. Work on personality [16] also suggests that the consistency of character is significant and this would relate to importance of the long-term perspective central to advanced empathy [18].

4 The Significance of Empathy

Developing a rich and sensitive understanding of every child requires considerable empathy on the part of the teacher. Research into empathy and into teaching and learning in the sixties and seventies explored the concept in considerable depth and linked empathy with effective teaching. David Aspy [1] conducted considerable research into humanising the teaching and learning process (when computers were not widely used!) and Carl Rogers [26], amongst others, highlighted the central nature of this quality not only in teaching but in all caring relationships. Unfashionable in the eighties and nineties it is currently undergoing a revival of interest. Empathy is widely associated with development and learning from intensely personal development during therapy to intellectual, spiritual, creative and also moral development [15]. Rogers makes a direct comparison between therapy and teaching in this extract:

“Just as the client finds that empathy provides a climate for learning more of himself, so the student finds himself in a climate for learning subject matter when he is in the presence of an understanding teacher” [26, pg 8]

From a therapist’s point of view their empathy is crucial for a patient’s personal development but a teacher’s role is different. Teachers are obliged both to discover a pupil’s existing skills or understanding in a particular subject area and extend them. In order to do this most effectively they have to know the child as a person as well as be aware of their academic understanding. They have to nurture their sense of self and support their academic success, which can also further develop their sense of self [24]. They may also develop their students awareness of other people, through simultaneously valuing them and

opening their eyes to other attitudes and understandings very different from their own. The summary of research evidence into empathy listed below, combines childrens' and teachers' opinions with evidence from observations in classrooms.

5 Recent Research into Empathy in Teaching and Learning

Research from an ongoing project (still in progress) into empathy in teacher/pupil relationships in which both pupils and empathic teachers were interviewed and observed in the classroom gives many pointers to the behaviours, attitudes and characteristics which might signify an empathic approach in the classroom [5,6]. The central aim of the study is to examine the role of empathy in teacher/pupil relationships and the role of these relationships in pupils' moral and academic development. Pupils and teachers were observed in the classroom and interviewed about their relationships and the role of empathy or understanding. There was considerable common ground agreed by both pupils and teachers about the characteristics and attitudes that denoted empathy and understanding and about the characteristics that denoted lack of understanding. Tables 1 and 2 represents a summary of the key characteristics identified so far, which may be helpful to AI in education.

6 General Constraints on Empathy in Classrooms

Though there is insufficient space in this paper to consider the constraints on the use of empathy in classrooms in any depth, it was clear that in lessons in which teachers were able to devote their sole concentration to single pupils, a much greater level of empathy was observable. The lack of time, the large numbers of pupils and the predominance of control in a normal classroom, as well as adherence to a rigid curriculum, seem to be some of the greatest inhibitors to a teacher displaying empathy and supporting effective individual learning.

7 Empathy in Intelligent Systems

Whilst some traits identified in existing intelligent systems could be identified in this research as being empathic for both pupils and teachers, some can be seen as unempathic. For a system, however intelligent, being human is not an option and a computer has distinct disadvantages in respect of voice tone, interpersonal elements, rigidity, ability to elicit understanding and to show and reflect real emotions. However it also has advantages, it can empower, give choice, be there when needed, give one to one attention, cater for different needs and so on. Building in the positive characteristics of empathy to our systems and eliminating the negative could well make them much more effective.

The teacher always has to be prepared to accept and support pupils, whatever level they are working at. There is no room for impatience and intolerance

Table 1. Characteristics of the Empathic Teacher pt 1

Empathic	Unempathic
Attitudes	
Open; warm; relaxed; good-humoured; Sees class as a group; not individuals; not fair; ensures fairness; models and expects interested in developing individuals; more common courtesy; explains how children interested in teaching subject; can be im-	patient; intolerant of some pupils' weak-
should work or behave in an under-	standing way rather than criticising their
present work or behaviour	nesses or even whole class; does not listen; finds it hard to change tack if lesson not going well and finds it hard to apologise if wrong
Facial characteristics	
Frequent smiles; lots of eye-contact; gen-	Robot-like; not expressive; do not show
erally positive demeanour; expressive face	emotions; facial expressions not in tune
which shows emotions and can switch with words;	
emotions quite quickly; tends to reflect	
student emotions but also leads and in-	
fluences them e.g. if the teacher wants to	
encourage thinking/reflecting, she models	
a thinking face	
Voice	
Positive; encouraging; expressive; clear di-	Unemotional; efficient; business-like; very
rections when necessary; supportive; var-	matter of fact; tone not in tune with
ied; reflects accurately the meaning of the words	
words	
Body-language	
Uses gesture; animated; tactile; moves	Wooden; unapproachable; distant; for-
around; uses body for emphasis and ex-	mal; not animated
planation	
Positioning	
Generally gets closer to child; less dis-	More formal; distanced from children;
tance; less formality and in a large class-	front of class; higher than children
room provides one to one support when	
possible; moves around quite a lot; sits	
down with pupils; lowers whole body, of-	
ten down below student's level.	

(shown by agents tapping a foot or drumming fingers), in a classroom where the teacher wants to foster self-esteem. We need the best of the peer, facilitator and tutor characteristics and more if we are to create not just ‘believable’ intelligent support in that it appears to be life-like, but systems which take more responsibility for children’s learning and attempt to meet diverse and long-term needs. Combining these with systems and hardware that support interaction rather than decreasing it also maximises the human support and the time available for

Table 2. Characteristics of the Empathic Teacher pt 2

Empathic	Unempathic
Responses	
Knows and uses student names frequently; listens carefully to students; standing; not individualised; overrides; ignores them sole concentration when possible; elicits understanding from them; helpful responses to children's attempts; echoes and affirms their comments; tries to give a positive response but asks them to elaborate or develop response if weak; prompts and helps them when necessary; constructs answerable questions to build success and self-confidence; frequent use of the 'cloze' technique ¹ to build confidence	Responds more to whole class; not individualised; overrides; ignores pupils' comments; negative or unhelpful responses to children's attempts; does not value or clarify comments; doesn't spend time explaining or developing response if weak; problematic issues
Content of teaching	
Frequently initiates a session with some aspect of topic that relates directly to teaching; child's own experience; personal interest, relate to children's interests and humour and discussion of non academic issues interspersed at appropriate moments through lesson; the personal used as a vehicle into the subject matter	Sticks to curriculum/subject; blanket aspect of topic that relates directly to teaching; little differentiation; does not relate to children's interests and humour and discussion of non academic issues interspersed at appropriate moments through lesson; the personal used as a vehicle into the subject matter
Method of teaching	
Varied teaching strategies; relaxed but rigorous; involves changes of pace and style; adaptable and flexible; sessions well-structured; individualised and personalised wherever possible; use of differentiation — matches task to child; explains problem issues; takes time over any issues; prepares individual material for children who need it	More rigid; lacks variety; lacks interpersonal level; elicits less from pupils; more instruction; less reflection/discussion; less reflection of pupils interest; emotions
Other features	
Uses humour; 'not like a teacher'; in touch with student's interests; form personal relationships with each child; considers informal significant; very aware of individual social and emotional aspects; puts time and effort into relationships; concerned with out-of-school life of child; maintain a long-term view of the child's well-being.	Behaves like a teacher; lacking in humour; shows false emotion; does not concern themselves with personal issues with children; tends to ignore emotional aspects of interaction; not particularly interested in the child beyond today's class.

¹ The cloze technique involves giving students the first part of an answer with a key part missing but which can be inferred in part from the context e.g. So the man put on his hat and coat and went out of the ...?

such support. We need to look to develop the best artificial models and create systems which maximise quality human support in their implementation.

8 Interpreting the Invisible and Creating Ambiance

The close interaction that goes on in empathic teaching relationships means that teachers react and respond at times to what is unseen or unsaid, or virtually unseen and unsaid or to changes in behaviour which rely upon previously held or perhaps unconsciously held knowledge of the child. The teacher might become aware that a child has not joined in a conversation about home, for example, or perhaps they seem suddenly reluctant to talk in a group situation. Small clues alert them to emotional change and are pointers to problems of a personal or academic nature, which may or may not have their origin in school. Sometimes a child may just look a little unkempt or stop smiling as readily. These observations can be registered almost unconsciously and gradually they surface and the teacher initiates an enquiry. Much of the finer details of visual appearance, body language and voice tone register almost subliminally in our brains [4]. Sometimes the teacher might respond to a cue without ever consciously realising why. Involved in hundreds of personal interactions every day, teachers may become increasingly expert and even less consciously aware, if they remain open to the signs emanating from the students in their care. As well as offering intelligent support, the NIMIS classroom aims to offer enough computer-based support to allow teachers to have more time to use their expert skills and create time for teachers to respond to needs that they identify. The deliberate design of intelligent systems that aim to meet needs, but which complement human expertise, should produce a powerful combination.

If we are to envisage whole classrooms rather than design single applications, the ambiance is also crucial to learning. Children need to feel generally secure, happy and to have their needs met if they are to thrive in any environment. Teachers with large classes of pupils have to establish an atmosphere of safety and a framework of commonly accepted rules but these have to be lightly handled, comfortable and relaxed, not authoritarian. Creating this type of ambiance demands considerable skill from the teacher and is much harder in some classes than others. Computer systems, which are sensitive to diversity and aim to meet needs, can contribute to a positive classroom climate.

9 Teacher Characteristics and Interventions during Story-Creation in NIMIS

To try and create a system which values the human skills involved in teaching and learning, the NIMIS project is using a participatory design methodology. From the outset the pupils and teachers have been involved in the design of the system and software via interviews, discussions, classroom observations and video recording. The NIMIS research team also includes ex-teachers and psychologists. A key aim of the NIMIS project is to build on existing practice and

to take into account the whole classroom and the dynamics of the real learning environment, simultaneously enhancing it.

During the initial analysis of the NIMIS classroom in Glusburn School near Leeds, the behaviour and comments of teachers during story writing, at times, demonstrated similar characteristics to those observed during the empathy project. In videos, class teachers could be seen demonstrating and modelling listening skills, direct intervention, reflection and positive affirmation. They modelled accepting facial and body language and interested voice tone, eliciting and elaborating ideas from the children, echoing and valuing their responses. When it was appropriate, they also took the responsibility for guiding and suggesting and for ensuring turn-taking and modelling and voicing rules of common courtesy.

Some of the constraints which teachers described in the empathy project were also clear in the videos in the NIMIS project. Time is the crucial constraint and it is clear, that 26 children need more support than any one teacher and even, on occasions three adults in the room could offer at any one time. Equally the constraints of organising and managing the learning activities of 26 5–6yr olds mean that children sometimes have to be given tasks which they can accomplish independently. In this sense they may be working easily within their bounds rather than being stretched and challenged by contingent teaching.

The number of children and the English National Curriculum methodology also means that individual support is more frequently relinquished for whole class or group support. In this sense the teacher's empathy, however sensitive, can be severely limited by the constraints of the classroom, the curriculum and sheer lack of time to interact personally with every child. A teacher can try to reduce the effect of this by employing 'group empathy' but according to the research, this is much easier with more homogeneous groups [7]. In NIMIS the system and software should be able to meet need more quickly with individual support at every workstation, whilst supporting both peer and teacher interaction both through classroom and furniture arrangements, electronic communication and software which supports shared story writing. Different tasks and levels of support will meet the needs of different children, but intrinsic to the NIMIS classroom is the continuing role of the class teacher and peer support alongside the software support.

10 Conclusions

Though intelligent systems may succeed in meeting some degree of individual need, the complexity of human interaction in teaching is incredibly difficult to model, given the way that teachers have to interpret the visible and invisible states of around thirty individuals at any time. They have to explore beyond the surface impression to sense the feelings and experiences beneath. In the NIMIS classroom normal human interaction and all the complexity it involves will be encouraged and facilitated but the characteristics of the system may also help to create and sustain an effective learning ambiance. Working with children on a personal level, intelligent systems may be able to offer a degree of support

more rapidly than the teacher who has an entire class of children to work with. They can offer targeted support on different levels and with a variety of methods in conjunction with a visible personality who affirms, suggests, guides, models listening behaviour and sole concentration, supports sharing and helping, steps in with support when it is needed but withdraws and encourages self-sufficiency when appropriate. This ‘empathy’ extends beyond the interactions to the nature of the software itself, which can cater for different tastes and attitudes of its users.

In this sense empathy should extend beyond the audio-visual affective aspects directly into the knowledge domain. This type of empathy which links the cognitive and the affective is very powerful and strong in terms of its ability to affect learning [11]. The child should feel secure, cared for and valued in its work because of the understanding displayed by the combined effect of a supportive system and teacher. This understanding will then be linked directly to concrete support for literacy development, with encouragement and affirmation for the work already completed and guidance for the next stage. This is the type of empathy offered by expert teachers in the classroom. It is both affective and cognitive and, in combination, more effective. It has a moral dimension in its attempts to meet need and include, rather than to encourage the survival of the fittest and exclude. Open and accepting relationships, built on trust, allow each party to discover and appreciate the views and understanding of the other more thoroughly. This understanding enables teachers to teach more effectively, basing their decisions on a richer information base and enables the child to understand the teacher so that richer dialogue can ensue. The interplay of feeling and understanding is central to learning as what we feel about our learning affects what we learn. A sense of security makes risk-taking possible. A feeling of acceptance and success allows acceptance of others and joint understanding to develop. Knowing support is there when needed means individuals are empowered, are able to experiment and feel secure enough to fail at times. The support required will be different for every individual. These feelings about learning are not peculiar to children but are common to adults too [8]. How open people are to expressing their feelings also depends on their relationship with their tutor. If they experience criticism or lack of interest, they may eventually refrain from expressing themselves unless they are very self-confident. If they feel dominated or made fun of, they may become introverted and fail to learn. If they consistently experience genuine concern for their development from their tutor, then they will feel able to make those leaps into the unknown which learning necessitates. A combination of sensitive computer-based support in the NIMIS classroom modelled on high quality, empathic teaching, combined with sensitive human support from teachers and peers in a positive learning ambience, may create an enhanced atmosphere in which learning can take place for all pupils.

Acknowledgements

This work is supported by the EU funded Networked Interactive Media in Classrooms (NIMIS) project no 29301.

References

1. Aspy, D. Towards a Technology for Humanising Education. Research Press, Champaign Illinois (1972). 22, 26
2. Bennet, N. and Dunne, E. How Children Learn — Implications for Practice. In Moon, B. and Shelton-Mayes, A. (eds.): Teaching and Learning in the Secondary School. Chapter 6. Routledge, London (1994). 22
3. Cassell, J. and Thórisson, K. R. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. Applied Artificial Intelligence 13 4/5 (1999) 519–538. 23
4. Claxton, G. Hare Brain Tortoise Mind. Fourth Estate Ltd, London (1997). 30
5. Cooper, B. L. Communicating values via the ‘hidden curriculum’ — messages from the teacher? In The Fourth Annual Conference on Education, Spirituality and the Whole Child. Roehampton Institute, London (1997). 27
6. Cooper, B. L. Exploring moral values — knowing me, knowing you ...aha! — rediscovering the personal in education. In The Fifth Annual Conference on Education, Spirituality and the Whole Child. Roehampton Institute, London (1998). 27
7. Cooper, B. L. Disingenuous teachers, disenchanted learners — in search of authenticity in teacher/pupil relationships. In The 4th Annual Conference of the European Learning Styles Information Network. University of Central Lancashire (1999). 31
8. Cousin, G. and Davidson, A. Information technology and the affective domain. In The 4th Annual Conference of the European Learning Styles Information Network. University of Central Lancashire (1999). 32
9. Goleman, D. Emotional Intelligence. Bloomsbury, London (1995). 22
10. Du Boulay, B., Luckin, R. and del Soldato, T. The plausibility problem: Human teaching tactics in the ‘hands’ of a machine. In Lajoie, S. P. and Vivet, M. (eds.): Artificial Intelligence in Education: Open Learning Environments: New Computational Technologies to Support Learning, Exploration and Collaboration. IOS, Amsterdam (1999) 225–232. 25
11. Fraser, V. The importance of story in children’s learning: The lessons of research. The Use of English, 50 1 (1998). 32
12. Gardner, H. The Unschooled Mind. Fontana, London (1993). 22
13. Goodman, B., Soller, A. L., Linton, F. and Gaimari, R. Encouraging student reflection and articulation using a learning companion. International Journal of Artificial Intelligence in Education. 9 (1998) 237–255. 24, 25
14. Hesten, S. The Construction of an Archive (on Heathcote). Unpublished PhD thesis, Lancaster University (1995). 22
15. Hoffman, M. L. Moral development. In Mussen, P. H. (ed.): Carmichael’s Manual of Child Psychology. Wiley, New York (1970). 26
16. Isbister, K. and Nass, C. Consistency of personality in interactive characters: Verbal cues, non-verbal cues and user characteristics. International Journal of Human Computer Studies (In Press). 26
17. Johnson, W. L., Rickel, J. W. and Lester, J. C. Animated pedagogical agents: Face to face interaction in interactive learning environments. International Journal of Artificial Intelligence in Education 11 (2000) To appear. 23
18. Kozeki, B. and Berghammer, R. The role of empathy in the motivational structure of school children. Personality & Individual Difference 13 2 (1992) 191–203. 26

19. Lepper, M. R. and Chabay, R. W. Socializing the intelligent tutor: Bringing empathy to computer tutors. In Mandl, H. and Lesgold, H. (eds.): *Learning Issues for Intelligent Tutoring Systems*. Springer, New York (1988) 242–257. 26
20. Department of Education and Science. *Discipline in Schools*. HMSO, London (1989). 22
21. Moon, B. and Shelton-Mayes, A. *Teaching and Learning in the Secondary School*. Routledge, London (1994).
22. National Commission on Education. *Success Against the Odds: Effective Schools in Disadvantaged Areas*. Routledge, London (1996). 22
23. Paiva, A., Machado, I. and Martinho, C. Enriching pedagogical agents with emotional behaviour — the case of Vincent. In *AIED'99 Workshop on Animated and Personified Pedagogical Agents*. Le Mans (1999) 47–55. 23
24. Purkey, W. W. *Self-concept and School Achievement*. Prentice-Hall (1970). 26
25. Robertson, J., Good, J. and Pain, H. BetterBlether: The design and evaluation of a discussion tool for education. *International Journal of Artificial Intelligence in Education* 9 (1998) 219–236. 24, 25
26. Rogers, C. R. Empathic: An unappreciated way of being. *The Counselling Psychologist* 5 2 (1975) 2–10. 22, 26
27. Vygotsky, L. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press, Cambridge, MA (1978). 22
28. Wood, D. *How Children Think and Learn. Understanding Children's Worlds*. Blackwell Publishers, Oxford (1988). 25
29. Zimmer, B. The empathy templates: A way to support collaborative learning. In Lockwood, F. (ed.): *Open and Distance Learning Today*. Chapter 14. Routledge, London (1995). 23, 25

The Communication of Meaningful Emotional Information for Children Interacting with Virtual Actors

Pat George and Malcolm McIlhagga

Interactive Media & Virtual Environment Lab,
School of Cognitive and Computing Sciences, The University of Sussex,
Falmer, Brighton, BN1 9QH. UK
{pgeorge,malcolm}@cogs.susx.ac.uk
<http://puppet.cogs.susx.ac.uk/>

Abstract. In a virtual environment where users can interact with agents, the way in which the agent displays its internal emotional state and what the user infers from this display is central to effective affective communication. Given the reliance on visual expression of emotion in current VR Systems, any differences between children and adults in the way such information is utilised and perceived raises interesting design issues in relation to the experience and age of the user. Here we discuss some of these issues in relation to the PUPPET project, where the goal is to provide a virtual play environment, using a theatre metaphor, to enable young children's (4-8 years old) story construction. This paper first presents a summary of previous research that demonstrates the special meaning that young children place on facial expressions when gathering information about emotion. Three empirical studies are presented that explore these issues using prototypes of our system. Some design solutions are suggested in conjunction with directions for future empirical work.

1 Introduction

Studies of affective communication with and between children have shown that several variables can change the way that this is achieved: particularly those of age and experience. In comparison to adults, there may be substantial differences for younger children in the way that they perceive affective expression and the cues that they rely on when making judgements about meaning. This is especially true with regard to visual signals.

These findings have relevance for the increasing number of attempts to design characters in virtual worlds which are capable of conveying some degree of meaningful and believable affective expression. There has been substantial progress with the design of agents for adult interaction, especially in e-business applications such as the Extempo gallery (<http://www.extempo.com/>) and for 'helpers' in

pedagogical applications e.g. the Intellimedia project at North Carolina State University. However, for young children, it is by no means clear how we should proceed since there is a lack of research guidance on how they represent (understand) virtual characters in general and, therefore, might use (interpret) affective signals in particular. Further, technical constraints in the speed and complexity with which many systems can produce real-time interactive visualisations of expressions, e.g. due to rendering limitations, means that we may need to work with simplifications from the real world. This is not necessarily a limitation, as we know from the effectiveness of cartoons, but the situation can be more complicated for interactive virtual systems, for example in terms of a very limited capacity for natural language processing (NLP). This means that there is an increased reliance on the visual interpretation of emotions, c.f. the likely situation in real life. Thus, it is an empirical question as to how system objectives might be realised in particular cases.

These issues are a central concern of our own research project, PUPPET, where the goal is to provide a virtual play environment, using a theatre metaphor, to enable young children (4-8 years old) to be more creative in their story development. The project aims to give the children the facilities to interact with characters, such as animals on a farm, allowing them the ability to choose one of the characters as their avatar. For the younger children, 4-5 year old, a major goal is to encourage the child to explore the environment, to discover and interact with the other characters (agents) within the virtual world. They are asked to observe and interpret the character's behaviour and they discover that the agents behave in different ways, following different kinds of rules, depending on their intentions and motivations. Clearly, given this aim, it is essential that we have some principled way of deciding how to visualise the affective state of agents—the most significant cue that these children have available to make their judgements. Specifically, we need to think about representation of facial expressions.

Below we describe how we have used a variety of exploratory studies in order to make and evaluate design decisions about the affective expression of the agents. First, we begin with a brief survey of the main conclusions of the psychological literature that are relevant to these issues.

1.1 Young Children's Ability to Recognise Emotion

Human emotional states can be categorised by: (i) the subjective experience of the emotion—happiness, sadness etc.; (ii) the physiological changes which occur—butterflies in the stomach, goosebumps, sweating etc.; (iii) the behaviour associated with a particular emotion—smiling, crying, frowning, running away, being frozen to the spot etc. From the point of view of the perceiver, any behaviour can be used to infer emotion—perhaps the most obvious would be something like crying, running, stamping of feet or blushing. However, recognising an emotion can depend upon the salience of many factors including the context of the behaviour. Apart from the context and the behaviour, adults tend to make complex inferences, which take into account the *beliefs* and *desires* of the individual concerned. Research specifically concerned with young children suggests that they may not normally utilise as many factors as adults when they gather emotional information.

Generally, young children tend to be more constrained (than adults) by what they can see (e.g. [31]; [26]). With specific reference to making inferences about emotion, [30] suggests younger children's understanding of emotion (between age 4 and 8 years) is primarily situation dependent. More specifically, Reichenbach and Masters [25] have demonstrated that when children are given a choice of cues to emotion, they tend to believe in the *veracity of facial expressions*. Indeed, Gnepp [14] and Iannotti [17] found that young children tend to rely on *facial expressions to read emotions even when they are in conflict with contextual cues*. This suggests that when children gather emotional information they tend to be particularly influenced (and biased) by facial expressions.

The research literature concerned with recognition of emotion also tends to place a heavy emphasis on the face, some researchers suggesting that a necessary prerequisite for the inference of emotion is *an ability to classify the facial expression of the person in question* (e.g. [3], [7] see [1] for a general introduction to the importance of faces for recognition of identity, age and gender, and facial expression of emotion). This is probably not surprising for human to human interaction given the attention on the face during social interaction and communication.

In an influential review of adult recognition of human facial expression, Ekman, Friesen, and Ellsworth [7], suggests that there is a high degree of consistency across empirical studies—all adults are able to distinguish between positive and negative emotional states, based upon the facial expression alone. However, the majority of adults can make much finer discriminations, which are thought to fall into seven distinct types of expression: *happiness, surprise, fear, sadness, anger, disgust, contempt, and interest*. These categories are often described as 'basic' emotions, which have come to be known as universal i.e. independent of culture.

From a developmental perspective, there is some argument about when and if children utilise all these categories. Some studies report an increasing ability in children up to 10 years of age [15], to discriminate between the types of emotional expression, whereas others using different tasks suggest that by 8 years old children use them all [28]. However, there is agreement that young children recognise positive emotions (joy or happiness) from a much earlier age than other emotions (e.g. [2]; [19]; [21]; [22]).

Using the photographs of human faces displaying the expressions identified by Ekman [5] (happy, sad, anger, fear, disgust and surprise), Markham and Wang [24] explored recognition of emotion for Chinese and Australian 4, 6, and 8-year old children. Although there were some differences depending upon the culture of the children, the developmental trend was clear—performance improved significantly between 4 and 6 years, with some further improvement between 6 and 8 years. Overall, their results suggest that children do find some facial expressions more difficult than others: younger children (between 4 and 6 years) tended to group some emotions together. While happy faces were easily differentiated, faces displaying fear were often classified as sad, and faces displaying disgust were often categorised as angry.

One limitation of many studies interested in developmental aspects of human face recognition is that they have tended to use the faces of adults as targets for the children to make judgements about. Given that judgements about faces displaying age-characteristics of a similar age to the perceiver (e.g. child faces) show a marked

improvement in recognition performance (see [13][12] for a discussion), it is possible that any developmental difference may have been overstated. However, whether this bias is constrained to human faces or is extended to cartoon like representations is an empirical question that deserves further consideration. Assuming that the own age-processing bias extends to more abstract representations of faces, this may be an important aspect of designing virtual characters and for designing cues for the children to recognise particular expression/emotion couplings.

Another factor is that children are also sensitive to the *degree of intensity of affect that the face displays* [24]. An interesting study by de Gelder, Teunini and Benson [10] produced ‘blends’ using morphing techniques of facial expression derived from four basic emotions (anger, fear, happiness and sadness). Children were asked to categorise three continuums that were linear transforms from one expression to another—anger/sad, happy/sad, angry/fear. The findings suggest that young children can judge whether a given facial expression is a ‘good’ or ‘poor’ example of the facial expression represented. Compared to adults, children seem to find faces displaying a happy expression the easiest to distinguish, with some confusion of discrimination with other emotions (fear with sadness, disgust with anger).

Related to degree of intensity, a large number of studies suggest that children (and adults) find *caricature faces* easier to process than normal faces (e.g. recognition of identity, age and emotional state is easier with caricatures of normal faces—see Ellis [8] for a discussion). The caricature type of distortion is thought to be meaningful for recognition purposes because it emphasises specific aspects that make a face distinctive, or different from others [17]. The important point here is that exaggerated facial expressions give more cues to the emotional state.

At this point it is important to point out a further, major methodological limitation of many research studies in that they have very often used *static* photographic images. These, necessarily, do not take into account the way a face may change or the speed with which the change takes place. However, there are some studies that have highlighted the more dynamic aspects of emotional expression. For example, [16] found that adult observers were able to categorise dynamic faces changing from neutral to surprise easily, but faces changing to happy were very difficult (best performance was for surprise, followed by disgust, fear, sadness, anger and then happiness). Humphreys *et al* [16] suggest that surprise is most easily perceived because of the facial movement from neutral (a sudden expansion in the lower part of the face which is caused by the dropping of the jaw, along with an expansion in the lower area of the eyes which caused by a raising of the eyebrows, and a corresponding compression of the forehead). Indeed, Ekman and Friesen [5][6] discussed a number of situations in which the rapidity with which components of an emotional expression are displayed carries information about the exact significance of the expression, and have come up with a facial action coding system (FACS).

Although these studies highlight the importance of dynamic facial information in conveying emotional expression for adults, there does not appear to be any obvious research with young children. However, it is certainly the case that dynamic changes convey information in the ‘real-world’ which has direct implications for the design of animated characters.

In summary, young children tend to rely on the emotion conveyed in facial expressions, even when they might be in conflict with the context [25], [14], [17].

These findings have high relevance for Puppet, suggesting the need for specific empirical research into the ways in which facial expression can be effective in a Virtual Environment. For example, *exaggeration of emotional expression*, by *type* (using caricatured faces) or in terms of the *temporal information* (rate and method of change from one emotion to another), provides a clear direction for communicating meaningful emotional information about the internal state of a virtual character. These points may be an effective design solution. Another factor concerns the age-characteristics of faces: faces with younger age-characteristics may be easier for young children to gather emotional information.

The review given above demonstrates that there are a wide variety of issues which are relevant to the way in which we might design the affective behaviour of virtual characters for young children. In PUPPET we have been concerned with establishing the ways in which expressions are perceived and utilised by 4 to 8 year olds. This paper now go on to briefly describe some of our initial exploratory empirical work into young children's abilities to select and respond to particular forms of facial expression.

2 Empirical Studies

There is a wealth of studies that need to be done in the present context. Here we describe three that address some of the design issues concerned with how children represent (understand) and how they interpret (use) emotional information, in isolation and in context.

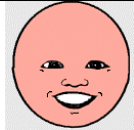

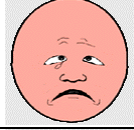
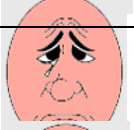
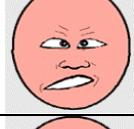
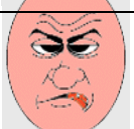
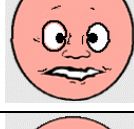

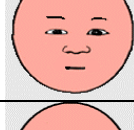
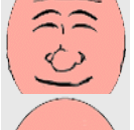
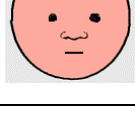

2.1 Study 1 - Discrimination of Cartoon-Like Facial Expressions

An important design issue for PUPPET is how to provide the child with a means of changing the emotional state of the agent/avatar. For example, the child may want it to express happiness, anger etc. after an interaction. Given the age of the children, a pictorial interface was considered the best solution. However, this 'emotional' interface needs to serve the contradictory function of allowing children to both recognise clear emotional categories, while at the same time giving them enough flexibility for creative and imaginative expression.

The research outlined in the previous section highlights the main facial expressions that are meaningful for children, in terms of their ability to discriminate between them (e.g. happy, sad, angry and frightened). For the purposes of our first prototype, two additional categories were selected: a neutral face which provides a position from which change could be determined; a calm face was also thought to be important since it would fit in with the dimensions identified by our partners who have developed a 'mind' for these characters (see Andre et al. in this volume "Integrating Models of Personality and Emotions into Lifelike Characters").

Empirical question: Can children discriminate between generated facial expressions?

Two sets of faces were produced, one set was derived from caricatured photographs of a young child modeling the expressions (Set 1). The second set was derived from cartoons of an older face also displaying caricatured expressions (Set 2). Sixteen children aged 5 and 6 years were seen at a local school. The children were presented with these faces and were asked to describe each of them in turn in their own words.

	Set 1	Set 2	Description of facial expression
Happy			"Happy": All children were able to categorise these faces as happy. Set 1 face produced more variation (e.g. happy, smiling, grinning, laughing) than Set 2 (only happy and smiling).
Sad			"Sad": Most children were able to categorise these faces into Sad, Crying, Grumpy or lonely. There was no distinction between the two sets.
Angry			"Angry": Set 1 produced more flexibility than set 2, although the majority of descriptions were similar (grumpy, angry, funny, mad, cross and scared).
Frightened			"Frightened": Some children were not sure about Set 1 face (e.g. funny, sick), but the majority of children described the faces consistently (e.g. scared, worried, frightened, shocked, terrified, spooky)
Calm			"Calm": These faces produced the most flexible descriptions. Set 1 faces included more diverse emotional categories (normal, a little angry, thinking, quite happy, tired) than Set 2 (happy, smiley, funny)
Neutral			

2.2 Overall Findings

Basic emotions (happy, sad, angry and frightened):

The faces displaying happy and sad expression seemed to be the easiest for the children to categorise - at least in terms of the descriptions that they gave. While the angry and frightened faces produced more flexible descriptions, the majority of

children were able to categorise the emotion in line with the intended display. There seemed to be little difference between the two sets although Set 2 produced more consistent descriptions than Set 1. Overall, these results are broadly similar to previous research using human faces, which suggests that both these sets of faces are 'good enough' representations of the expressions.

More neutral faces (calm and neutral):

These faces produced the most flexible descriptions from the children with some children suggesting that they might display mixed emotions. In terms of flexibility, Set 1 was better than Set 2. Given that a major aim of Puppet is for children to develop creative and imaginative stories Set 1 may be a more suitable set of faces to incorporate into the design of a 2D pictorial user interface.

2.3 Study 2(a) - Building Up Characters – Linking Emotion to Action and Intention

This study was designed to explore how children incorporate or utilise emotion when building-up a character. Five 4-year old children were seen for this purpose.

Empirical question: How easily do children map facial expressions onto a 'thinking' and 'acting' character?

The children were asked to build up a character by choosing legs, arms, and faces after being given a stem of a story (e.g. "imagine you are a farmer and you have lost a lamb..."). In one window a stick figure was displayed and in another window a set of options or attributes. Initially the figure was motionless and inactive. First, the child could select an action attribute such as chasing, talking, and searching. When an action was selected the child could modify the action by selecting from such categories as climbing shouting, walking. Finally, the child could attribute the character with an emotion (see figure 2.2 below). For a complete picture of the combination of actions, modifiers and emotions (see Figure 2.1)

Once the child had attributed the stickman with an emotion, action or action modifier the stickman in the first window was animated to give direct feedback to the child.

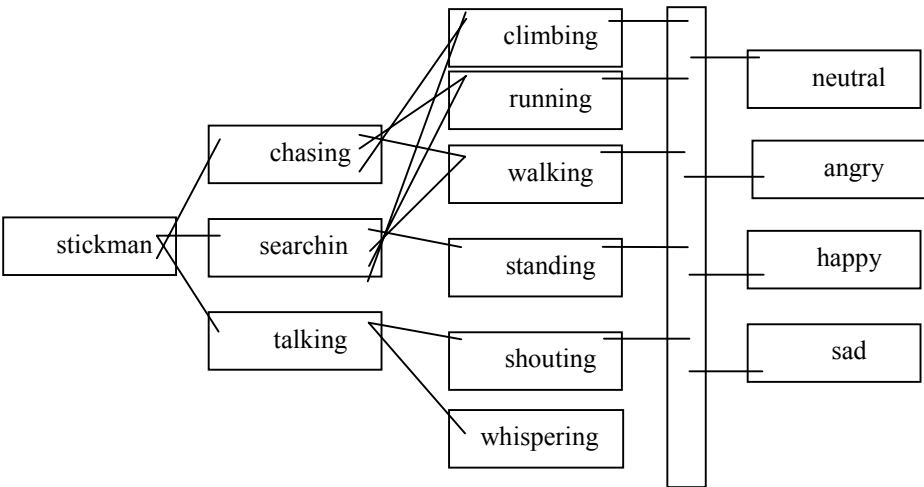


Figure 2.1 Possible combinations of attributes for stick man

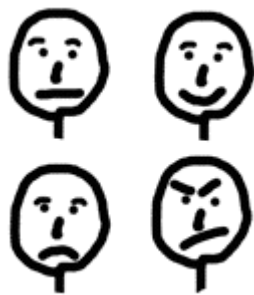


Figure 2.2: Faces used for the stick man-displaying a neutral, happy, sad and angry expression

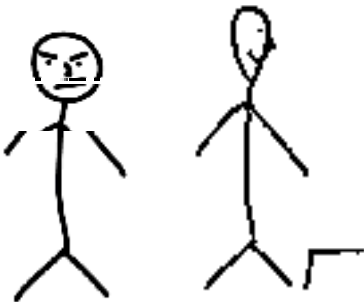


Figure 2.3: These images are snapshots taken from animations built-up by the child selecting different combinations. The left is a talking (mouth moving in the animation) angry, walking (walking movement in the animation) man. The right is a chasing (moving quickly) climbing (stairs to the far right), happy, man.

When the animation was running, the children were encouraged to elaborate what the character was thinking, what they were doing, as well as what they were feeling if they had not already done so.

2.4 Findings

Overall, although the children were able to build up the characters in this fashion, and use facial expressions that were appropriate (e.g. if they reported the farmer was angry they would pick an angry face), they did not find this task engaging. This may

highlight one of the problems in isolating one specific design aspect of the proposed system and building a prototype program to investigate possible design issues, i.e. attributing actions and facial expressions without providing a framework (e.g. appropriate or supportive context).

2.5 Study 2(b) - Building Up Scenarios – Emotions in Context

An interactive prototype within the context of a farmyard was developed to explore how far children can drag'n'drop appropriate behavioural attributes, from a palette, to the characters as a function of a simple story-line. Part of this involved their selection of facial expressions.

Empirical question: How easily do children map facial expressions within the context of their developing story-line

The prototype had various functionalities (1) animal sound buttons (which played pig, cow and horse sounds), (2) move buttons (which made the animals move individually or in combination from one side of the river to their house), (3) facial expression buttons (which displayed a happy, thoughtful and an angry expression), and (4) farmer voice buttons (which played different sentences designed to match happy, thoughtful, and angry).

Figure 2.4 below shows the full set of buttons that were available to the children, and a specific scenario that was created by one child. However, the different buttons only appeared at specific times. For example, the first set was the animal sound buttons. When the children attached sounds to the animals, they would disappear and the movement buttons were shown. Only if the child selected a 'fast' cow and a 'fast' pig and they were sent home together, would a collision appear (as was the case in Figure 2.4). At this point, a farmer would walk onto the stage and the face buttons would appear, when one was selected it changed the farmer's face appropriately, then the voice buttons would appear.



Figure 2.4 Prototype with all buttons displayed

2.6 Findings

The children were able to drag and drop and mouse click with little practice. The sounds were particularly enjoyable. The function and number of buttons is very important- the simpler (and more direct) the better. Although young children needed some support in their exploration, they were able to associate facial expression in this context.

2.7 Study 3 - Spontaneous Story Generation with Cartoon Faces

This study was designed to explore how and if young children will use facial expressions within their stories.

Empirical question: How children use facial expression for short story generation

Sixteen 5 and 6 year old children were asked to generate a short story from the following stem, "Can you tell a story about a farmer, and a pig that has ran off...", no other details or instructions were given. They were provided with two pictures - one displayed a farmer and another a pig- both with neutral facial expressions (see Figure 3.1). The children were also provided with facial expressions (one set for each character) and told that they could select from these facial expressions and attach them to the characters as and when they wanted to (see Figure 1 for the faces).

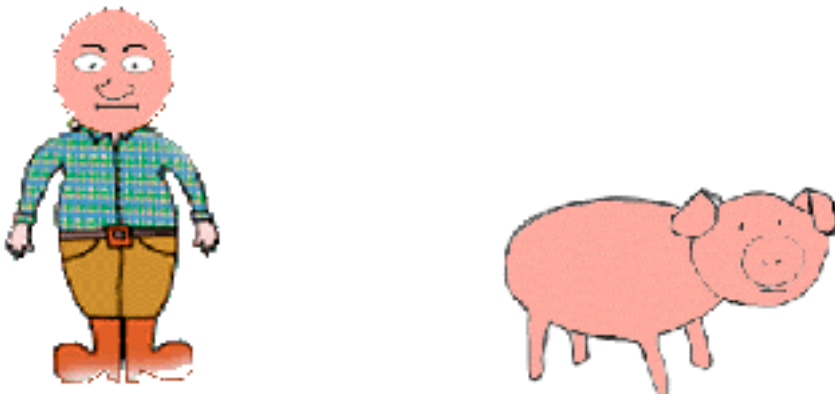


Figure 3.1 The two images used as a starting point

2.8 Findings

All the children responded with enthusiasm and seemed to enjoy this task very much. Although they were encouraged to tell a short story the average number of story-lines was five (range 3 to 9). The average number of faces selected was also 5 (range 2 to 8) and the average number of different emotional states reported for both characters was 3 (range 1 to 4).

A rather surprising finding was that all the children seemed to adopt a similar strategy of first selecting a face for the characters, and after that, elaboration the characters actions, intentions and emotional state (e.g. an angry face was selected for the farmer and a happy face was selected for the pig, then the child reported what the farmer and the pig were doing and then what they felt - see Figure 3.2).

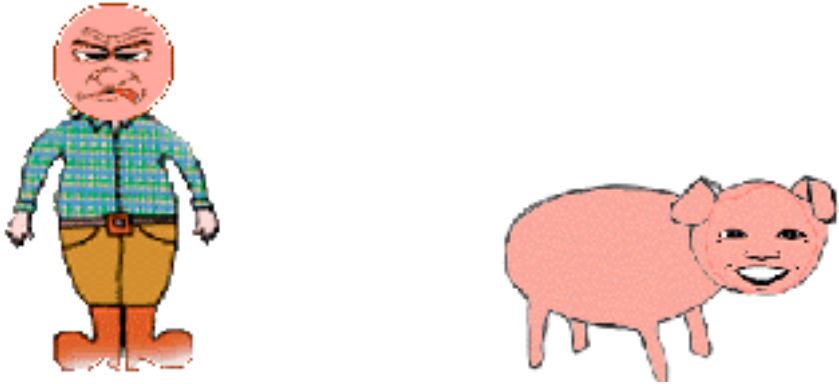


Figure 3.2 A selection from one child's story, "The farmer is angry because the pig has ran off. The pig is happy because he can explore...."

There are a number of issues raised by this task, that are empirical questions in their own right and deserve further investigation. An important point is that the children did seem to enjoy the task very much. The results do suggest that using faces may in some way allow the children to be creative and/or support the expression of story-lines that include the intentions and mood of the character. However, an empirical study specifically addressing this point is indicated.

2.9 Overall Summary of Findings

In the following studies we have addressed several issues. Children were able to discriminate between the different cartoon-like faces; previous research describes similar discriminations using human faces. Children were able to build up a character actions, faces and action modifiers, and were able to use facial expressions within a story telling context. The final study shows that children can use facial expressions creatively: facial expressions can be used to elaborate the protagonists intentions, goals and emotion.

3 Conclusion

Children are primarily influenced by visual cues, particularly facial information. What sort of faces should we use? For adults and children, faces that are caricatures are easier to process and convey more information. Faces that reflect the age of the user can also be used to convey more information. Familiarity is also a factor: if the user

has interacted with an someone over an extended period, emotional information can be conveyed in a more subtle fashion. Although this is known for human faces, this still has to be investigated for cartoon-like faces.

When thinking about how to approach design decisions for our project, there is a marked lack of guidance in the literature as to how young children might understand and interpret the behaviour of virtual characters and their emotional signals. Our findings indicate that children are able to use facial expression when building up characters in specific situations. They can also use facial expression in the context of storytelling. It appears from our studies that using facial expressions may *even* better enable children to elaborate on a character's intentions, goals and emotion. This issue is a matter for further investigation.

In our proposed system we need to provide a means of enabling children to change the emotional state of a character. Previous research has provided little direction as to an appropriate approach to this problem. Given the age of the children and the different types of characters that will inhabit the world, abstract cartoon like representations provide flexibility. They need to do so in order to be used for multiple characters. Previous research that has used human faces has provided as starting point as to which basic categories children can discriminate easily. We have extended this research by using cartoon like representations. Our results suggest that although the children can discriminate the cartoon like representations easily, they also need to have faces that are 'neutral' (not just basic emotions as identified by previous research) as they support creative and imaginative play. Our final study demonstrated that emotional expressions that are flexible can be used in a creative and imaginative way by the children in an improvisational situation. Thus, cartoon faces that are derived from exaggerated versions (caricatured) of child faces provided a design solution for our system.

We suggest that in order to empower the user with a richer model of the characters that they interact with, the designer should focus their attention on mechanisms, such as facial expression and exaggeration, which better communicate meaningful emotional information.

Acknowledgements

The authors gratefully acknowledge support from the EU Espirit i3-ESE programme, Project EP 29335, and especially to Dr. Mike Scaife and Dr. Yvonne Rogers, who are the principle investigators on the project. We also would like to acknowledge the contribution of our partners: LIA, University of Aalborg, Denmark, IDAU, University of Aarhus, Denmark and DFKI, Saarbrücken, Germany. Finally, we would like to thank all the children who took part in our studies, their parents and the schools in East Sussex, UK.

References

1. Bruce, V.: *Recognising Faces*. Lawrence Erlbaum Associates, London (1988)
2. Camras, L. A., and Allison, K. Children's understanding of emotional facial expressions and verbal labels. *Journal of Nonverbal Behavior*, 8, (1985), 15-38.
3. Cole, J. *About Face*. MIT Press, Cambridge, MA. (1998)
4. Elliot, C.: Research Problems in the Use of a Shallow Artificial Intelligence Model of Personality and Emotion Proceedings of the 12th National Conference on Artificial Intelligence. AAAI Press **1** (1994) 9-15
5. Ekman, P.: *Pictures of facial affect*. Consulting Psychologists Press, Palo Alto, California (1976)
6. Ekman, P. and Friesen, M. V. Facial action coding system: a technique for the measurement of facial movement. Consulting Psychologists Press, Palo Alto, California. (1978)
7. Ekman, P., Friesen, W.V., Ellsworth, P.: What are the similarities and differences in facial behaviour across cultures? In: P. Ekman (Editor) *Emotion in the human face*. 2nd Edition. Cambridge University Press, Cambridge (1982) 128-144
8. Ellis, H. D.: Developmental trends in face recognition. *The Psychologist: Bulletin of the British Psychological Society* **3** (1990) 114-119
9. Fleming, B., Dobbs D.: *Animating Facial Features and Expressions*. Charles River Media, Inc. Rockland, Massachusetts (1999)
10. de Gelder, Teunisse, Benson: Categorical perception of facial expressions: categories and their internal structure. *Cognition and Emotion* **11:1** (1997) 1-23
11. George, P.A., Hole, G.J.: Recognising the ageing face: the role of age in face processing. *Perception* **27** (1998) 112-134
12. George, P.A., Hole, G.J.: Factors influencing the accuracy of age-estimates of unfamiliar faces. *Perception* **24** (1995) 1059-1073
13. George, P.A., Scaife, M., Hole, G. J.: Factors influencing young children's ability to discriminate unfamiliar faces by age. *International Journal of Behavioural Development*. In press. (2000)
14. Gnepp J.: Children's social sensitivity: inferring emotions from conflicting cues. *Developmental Psychology* **19** (1983) 805-814
15. Harrigan, I. The effects of task order on children's identification of facial expressions. *Motivation and Emotion*, 8, (1984), 157-169.
16. Humphreys, G. W., Donnelly, N., Riddoch, M. J. Expression is computed separately from facial identity, and it is computed separately for moving and static faces: Neuropsychological evidence. *Neuropsychologia*, 31:2, (1993), 173-181
17. Iannotti, R.J.: Effect of role-taking experiences on role taking, empathy, altruism and aggression. *Developmental Psychology* **14** (1978) 119-124
18. Leslie, A.M.: Pretence and representation: the origins of "theory of mind". *Psychological Review* **94** (1987) 412-422

19. Markham and Adams, The effect of type of task on children's identification of facial expressions. *Journal of Nonverbal Behavior*, 16, (1992), 21-39
20. Markham, R., and Wang, L. Recognition of emotion by Chinese and Australian children. *Journal of Cross-Cultural Psychology*, 27:5, (1996), 616-643
21. Michalson, L. and Lewis, M. What do children know about emotions and when do they know it? In M. Lewis and C. Saarni (Editors) *The socialization of emotions*, (1985), 117-139. Plenum Press, New York
22. Pitcairn, T. Origins and processing of facial expression. In A. Young and H. Ellis (Editors) *Handbook of research on face processing*, Elsevier, Amsterdam. (1989), 71-76
23. Quasthoff, U.M.: An interactive approach to narrative development. In Bamberg, M. (Editor) *Narrative Development: Six Approaches*. Lawrence Erlbaum, London (1997)
24. Markham, R., Wang, L.: Recognition of emotion by Chinese and Australian children. *Journal of Cross-Cultural Psychology* **27:5** (1996) 616-643
25. Reichenbach, I., Masters, J.: Children's use of expressive and contextual cues in judgements of emotion. *Child Development* **54** (1983) 102-141
26. Singer, J. L. Imaginative play and adaptive development. In J. H. Goldstein (Editor) *Toys, Play and Child Development*. Cambridge University Press, Cambridge. (1994)
27. Stevenage, S. Expertise and the caricature advantage. In T. Valentine (Editor) *Cognitive and Computational Aspects of Face Recognition: Exploration in face space*. Routledge, London (1995)
28. Tremblay, C., Kirouac, G., and Dore, F. The recognition of adults' and children's facial expressions of emotion. *Journal of Psychology*, 121, (1987), 341-350.
29. Valentine, T. A unified account of the effects of distinctiveness, inversion and race on face recognition. *Quarterly Journal of Experimental Psychology* **43:A** (1991) 161-204
30. Vinden, P.G. Children's understanding of mind and emotion: A multi-culture study. *Cognition and Emotion* **13:1** (1999) 19-48
31. Vygotsky, L. S. The role of play in development. In M. Cole, V. John-Steiner, S. Scribner, and E. Souberman (Editors) *Mind in Society*. Harvard University Press, Cambridge, MA. (1978)

Emotion and Facial Expression

Thomas Wehrle and Susanne Kaiser

University of Geneva, Department of Psychology,
40, Boulevard du Pont d'Arve, CH-1205 Geneva, Switzerland
Thomas.Wehrle@pse.unige.ch

Abstract. Facial expression is usually synthesized or predicted on the basis of a given emotion. The prototypical expressions for basic emotions (happiness, sadness, surprise, disgust, anger, and fear) as postulated by discrete emotion psychologists are rather consistently produced and interpreted among different cultures, and can be used as icons to represent a basic emotion. However, these facial expressions are actually rarely observable in every day affective episodes, because of the complexity and multi-functional nature of facial expression and a high inter-individual variance. This article sketches the different functions of facial expression and presents an alternative way to predict, synthesize, and understand facial behavior. We present an empirical setting, which allows us to collect behavioral data in human-computer interactions, and show some examples from empirical studies as well as a computational model for facial expression on the basis of appraisal theory.

1 Introduction

In recent years the importance of emotions and emotion expression in human interactions has been widely accepted. Studying the functions of facial expressions is not only of interest in the domain of psychology and other social sciences but also in the domain of computer science and computational intelligence. There exist an increasing number of research groups that are developing computer interfaces with synthetic facial displays (e.g., [1], [8], [33]). These researchers attempt to use these facial displays as a new modality that should make the interaction more efficient, while lessening the cognitive load. In addition, several researchers point out that automatic interpretation of gestures and facial expressions would also improve man-machine interaction (e.g., [9], [22]). This article aims at illustrating how an appraisal based approach to the understanding of the relation between emotion and facial expressions might be instrumental for these two domains and their possible applications, i.e., a) facial expression synthesis (animated intelligent agents), and b) automatic expression recognition (decoding the emotional state of the user).

The new field of "affective computing" located between psychology, engineering, and natural sciences demonstrates the promise for interdisciplinary collaboration in these domains [24]. Although the attempt to implement specific models of the emotion process is not new (for reviews see [23], [36], [44], [45]), the availability of powerful techniques in artificial intelligence and the increasing focus on enhanced user interfaces, which should improve the usability of computer applications render this approach particularly promising. This book, which is the result of an interdisciplinary workshop on affective interaction also demonstrates this progress. Within this field, an important distinction has to be made between theory modeling on the one hand, and artificial emotions, on the other. Whereas the latter domain is concerned with the application of emotion theories to user interfaces, and the design of autonomous emotional agents, theory modeling serves the purpose to improve our understanding of the phenomenon of emotion in humans (for more details see [42], [45]). Although our research interests are more linked to the latter purpose, we think that an interdisciplinary approach is useful and also necessary in order to develop efficient "emotional interfaces".

We are using human-computer interactions and interactive computer games in order to study the ongoing dynamic cognitive and emotional processes including the situated behavior and the verbal and nonverbal expression. The experimental setting allows us to study the dynamics of emotional episodes in a more interactive manner than it is usually done in classical experimental settings¹. One reason why traditional emotion induction techniques often trigger only rather weak and not clearly defined emotional states might be that subjects are not really involved and/or that they have to verbalize their emotional state instead of reacting to an emotion eliciting event or situation. This is diametrically opposed to one of the most adaptive functions of our emotionality, i.e., to guide our behavior in situations that are important for our goals and needs and that require immediate cognitive and behavioral responses. In real life situations, we are not capable of developing "the perfect solution" to a problem. We cannot consider every aspect of imaginable decisions or reactions, e.g., all possible short- and long-term consequences. Yet, in many cases, emotions help us to find good solutions. We refer to this as *emotional problem solving*. Emotional problem solving is not a yes or no decision making but a process that unfolds in emotional episodes. Decisions are adapted and changed according to the dynamic changes in the external environment and according to changes caused by internal processes, concerning for instance memory, motives, and values. Following cognitive emotion theory, we can describe an emotional episode as a process of *primary appraisal* (the subjectively estimated significance of an event for one's well being), *secondary appraisal* (the subjectively estimated ability to cope with the consequences of an event), *coping*, and *reappraisal* in a *transactional* interaction [20].

As suggested by Leventhal and Scherer [20], these appraisal processes can occur at different levels and are often very fast and automatic. Appraisal processes occurring

¹ One traditional approach to study emotions is by asking people to remember as vividly as possible a situation where they experienced a certain emotion. Another approach tries to evoke emotions with the aid of selected video sequences that are judged as emotionally arousing.

on the sensory-motor or schematic level are rarely or only with great difficulty accessible through verbalization. One reason for analyzing facial expressions in emotional interactions is the hope that these processes might be accessible or indicated by facial expressions. Another reason for analyzing facial expressions in experimental emotion research is that these are naturally accompanying an emotional episode, whereas asking subjects about their feelings interrupts and changes the ongoing process.

Starting from a description of the multiple functions of facial behavior, this article describes the methodological and theoretical challenges to studying the relation between emotion and facial expressions. We present two mainstreams in emotion theory, which is discrete emotion theory postulating *basic emotions*, and componential appraisal theory presenting an appraisal-based approach as an alternative conceptualization. Specifically, we discuss how conceptualizing facial expressions as indicators of appraisals instead of basic emotions might help to analyze and synthesize facial behavior in emotional interactions. Finally, we present an empirical paradigm for analyzing facial expressions in emotional interactions and discuss some results with respect to current emotion theory.

2 The Multi-functionality of Facial Behavior

Facial expressions can have different functions and meanings [5], [10], [27]. A smile or a frown, for instance, can have different meanings. It can be:

- *a speech-regulation signal (regulator)*: a listener response (back-channel signal), telling the speaker that he can go on talking and that his words have been understood.
- *a speech-related signal (illustrator)*: a speaker can raise his eyebrows in order to lay particular emphasis on his or her argumentation. The facial signals can also modify or even contradict the verbal messages, e.g., a smile that indicates that what is being said is not meant to be taken seriously.
- *a means for signaling relationship*: installing, maintaining, or aborting a relationship, e.g., when a couple is discussing a controversial topic, a smile can indicate that although they disagree on the topic there is no "danger" for the relationship.
- *an indicator for cognitive processes*: e.g., frowning often occurs when somebody does some hard thinking while concentrated attending to a problem, or when a difficulty is encountered in a task.
- *an indicator for an emotion (affect display)*: a person smiles because he or she is happy. Besides, affect displays that occur during an interaction can refer to the interaction partner (e.g., becoming angry with the other) but it can also refer to other persons or themes the interaction partners are talking about (e.g., sharing the anger about something).

When considering spontaneous interactions, it is very difficult to identify whether a facial expression is an indicator of an emotion (affect display) or whether it is a

communicative signal. To make things even more complicated, a facial expression can have several meanings at the same time: e.g., a frown can indicate that the listener does not understand what the speaker is talking about (cognitive difficulty); at the same time this frown is a listener response (communicative), signaling that the speaker has to explain his argument more appropriately; finally, it can indicate that the listener is becoming more and more angry (emotional) about this difficulty in understanding him, about the content, or about the way this interaction develops.

Considering the complex and multiple functions of facial expressions, research paradigms for studying emotions and facial expressions should fulfill the following requirements:

1. We need approaches to measure facial expressions objectively and on a micro-analytic level. The *Facial Action Coding System* (FACS [6]) lends itself to this purpose. FACS allows the reliable coding of any facial action in terms of the smallest visible unit of muscular activity (*action units*), each referred to by a numerical code. As a consequence, coding is independent of prior assumptions about prototypical emotion expressions. Using FACS, we can test different hypotheses about linking facial expression to emotions.
2. The current, concrete meaning of a facial expression can only be interpreted within the whole temporal and situational context. In everyday interactions, we know the context and we can use all information that is available to interpret the facial expression of another person. Therefore, facial expressions and emotions should be studied in an interactive context.

3 Contemporary Emotion Theories

Most contemporary emotion theorists (e.g. [4], [11], [13], [28]) consider emotion as a phylogenetically continuous mechanism for flexible adaptation, serving the dual purpose of rapid preparation of appropriate responses to events and of providing opportunities for re-evaluation and communication of intent in the interest of response optimization. Following Darwin's [3] view of expressions as rudiments of adaptive behavior which have acquired important signaling characteristics, a functional approach to the study of emotion presumes that motor expressions are both reliable external manifestations of internal affective arousal and social signals in the service of interpersonal affect regulation (see also [14], [2]).

There is a long tradition in emotion psychology of examining facial expressions as an observable indicator of unobservable emotional processes. Among emotion theories proposing implicit or explicit predictions for emotion-specific facial expression patterns, two positions can be distinguished. The first approach is situated in the tradition of *discrete emotion theories* and is represented by Ekman, Izard, and their respective collaborators (e.g., [4], [13]). The second approach has been suggested in the context of *appraisal theories of emotion* (e.g., [17], [26], [28], [29], [31], [46]).

3.1 Discrete Emotion Theory and Facial Expression: Basic Emotions

Discrete emotion theorists have studied facial expression as the "via regia" to emotions for many years. Most of their research concerning the universality of the so-called *basic emotions* is based on studies about facial expressions. These theories claim that there are only a limited number of fundamental or basic emotions and that for each of them there exists a prototypical, innate, and universal expression pattern. In this tradition, a process of blending or mixing the basic expression patterns explains the variability of emotion expressions commonly observed. According to Ekman [4], an interpretation of facial expressions must rely on the postulated configurations and not on single facial actions.

There is considerable evidence (reviewed in [7]) indicating distinct prototypical facial signals that can be reliably recognized across a variety of cultures as corresponding to the emotions of happiness, sadness, surprise, disgust, anger, and fear.

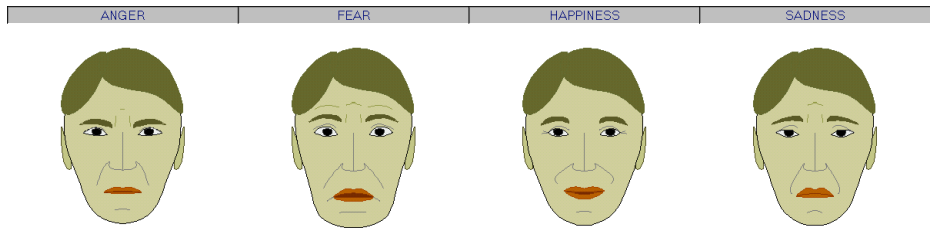


Fig. 1. Prototypical expressions as postulated by discrete emotion theorists for the emotions of anger, fear, happiness, and sadness

These patterns have been found in studies using photographs of posed facial expressions (Figure 1 shows some examples of basic emotion expressions synthesized with FACE [39]). However, these findings have not enabled researchers to interpret facial expressions as unambiguous indicators of emotions in spontaneous interactions. The task of analyzing the ongoing facial behavior in dynamically changing emotional episodes is obviously more complex than linking a static emotional expression to a verbal label.

Another problem not solved by the studies on the universal recognition of basic emotions concerns the dynamics of facial expressions. Generally, our theoretical knowledge of the temporal unfolding of facial expressions is quite limited. For example, one of the consistent findings in emotion recognition studies using static stimuli is that fear expressions are often confused with surprise. One explanation for this might be that the main difference between these two emotions resides in the respective temporal structure of the innervations in the facial musculature, which is not observable in still photographs. Another example of the relevance of temporal aspects is the fact that observers are very sensitive to false timing of facial expressions (e.g., abrupt endings or beginnings of a smile) when evaluating the truthfulness or deceitfulness of an emotional display. Although the importance of "correct" timing is widely accepted at a theoretical or phenomenological level, only a small number of studies have investigated this aspect systematically. The quantitative aspects of

spontaneous facial expressions deserve further investigation, especially with regard to the duration of onset, apex, and offset.

The limitations of a basic emotion approach also have consequences for affective human-computer interactions and the above-mentioned goals of expression synthesis and expression decoding. Those prototypical full-face expressions might serve as icons representing basic emotions in simple emotional agents (synthesis) or as stereotyped signals of the user to indicate “anger” or “joy” (automatic expression recognition). Such signals could be used instead of more complicated and time-consuming messages via the keyboard like “I do not understand” or “this is fine with me”. However, a basic emotion approach does not allow us to make inferences about the emotional state of a user when interacting with a computer. In the following, we discuss how conceptualizing facial expressions as indicators of appraisals instead of basic emotions might help to analyze facial behavior in emotional interactions and to develop more sophisticated synthetic agents.

3.2 Componential Appraisal Theory and Facial Expression

Appraisal theorists following a *componential approach*, as proposed by [11], [26], [28], or [31], share the assumption that a) emotions are elicited by a cognitive evaluation (*appraisal*) of antecedent situations and events and that b) the patterning of the reactions in the different response components (physiology, expression, action tendencies, feeling) is determined by the outcome of this evaluation process. For these appraisal theorists the complexity and variability of different emotional feelings can be explained without resorting to a notion of basic emotions. They argue that there are a large number of highly differentiated emotional states, of which the current emotion labels capture only clusters or central tendencies of regularly recurring ones, referred to as *modal* emotions. In line with this reasoning, facial expressions are analyzed as indicators of appraisal processes in addition to or as an alternative to verbal report measures. Facial expressions are not seen as the “readout” of motor programs but as indicators of mental states and evaluation processes. In contrast to discrete emotion theorists, they claim that single components of facial patterns do have a meaning and that this meaning can be explained as manifestations of specific appraisal outcomes.

Several appraisal theorists have made concrete suggestions concerning possible links between specific appraisal dimensions and specific facial actions [12], [14], [17], [32], [46]. Using the most recent version of FACS, Wehrle, Kaiser, Schmidt, and Scherer [46] have extended and refined Scherer’s original predictions linking facial actions to the postulated appraisal checks [29]. Scherer posits relatively few basic criteria and assumes sequential processing of these criteria in the appraisal process. The major “stimulus evaluation checks” (SECs) can be categorized into five major classes: 1) the novelty or familiarity of an event, 2) the intrinsic pleasantness of objects or events, 3) the significance of the event for the individual’s needs or goals, 4) the individual’s ability to influence or cope with the consequences of the event, including the evaluation of who caused the event (agency), and 5) the compatibility of the event with social or personal standards, norms, or values. As an example, Figure 2

shows the postulated facial expressions (action units) representing the appraisal profile for hot anger.

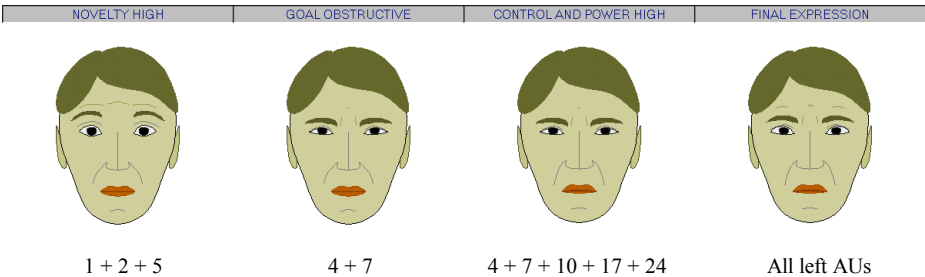


Fig. 2. Predictions for the appraisal patterns and the related facial actions for *hot anger* as published in [46]. From left to right, the pictures illustrate the sequential cumulation of appraisal-specific facial action unit combinations resulting in a final pattern, as postulated by Scherer [29]. Action unit numbers and names: 1 (inner brow raiser), 2 (outer brow raiser), 4 (brow lowerer), 5 (upper lid raiser), 7 (lid tightener), 10 (upper lip raiser), 17 (chin raiser), 24 (lip press)

4 Facial Expression and Emotion in Human-Computer Interactions

In spontaneous interactions, facial behavior accompanies emotional episodes as they unfold, and changes in facial configurations can occur very rapidly. In the preceding sections, we have illustrated some of the methodological and theoretical problems we are confronted with if we want to study the process of emotional interactions and its reflection in facial activity in interactive settings. To tackle some of these problems we are using a theory based experimental paradigm including computerized data collection and data analysis instruments on the one hand, and computer simulation and theory based synthetic stimuli on the other hand.

4.1 Experimental Setup, Data Collection, Data Analysis, and Modeling

In the following, we present some results from our research, using human-computer interactions for studying the dynamics and the interactive nature of emotional episodes. For this purpose, we developed the *Geneva Appraisal Manipulation Environment* (GAME [40]), a tool for generating experimental computer games that translate psychological postulates into specific micro-world scenarios (for details about theoretical and technical embedding of GAME see [16], [17], [19]). GAME allows automatic data registration of the dynamic game progress and the subjects' actions and automatic questionnaires. For example, participants' evaluations of specific situations are assessed by means of pop-up screens (which appeared after the completion of each game level) corresponding to 18 questions referring to Scherer's appraisal dimensions (SECs; see also [30]). While playing the experimental game,

participants are videotaped and these tape recordings allow an automatic analysis of the participant's facial behavior with the *Facial Expression Analysis Tool* (FEAT; [15], [34]). These facial data can be automatically matched with the corresponding game data (using the vertical time code as a reference for both kinds of data). In this way, the computer game provides a relatively small but complete context for the interpretation of the internal emotional and cognitive regulatory processes. This is an important point, because what is felt by the subject and what a certain facial expression means is often very context specific.

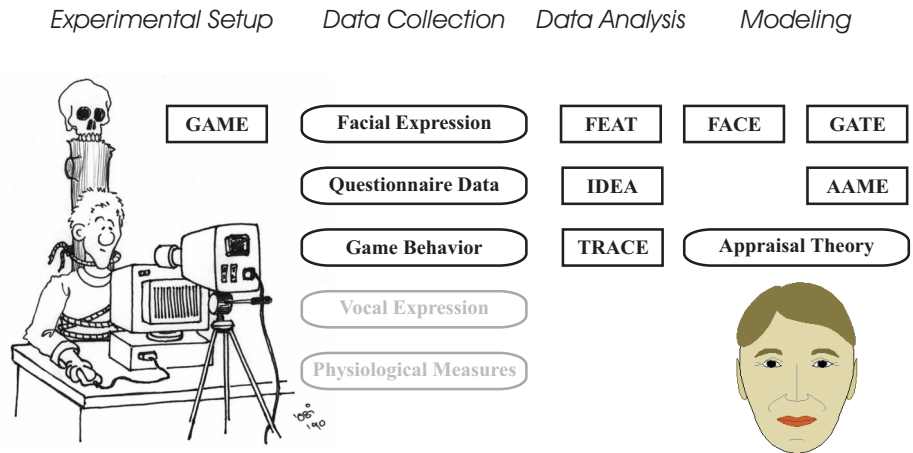


Fig. 3. The components of the experimental setup (GAME), including computerized tools for automatic data collection and data analysis, as well as tools for synthetic modeling and simulations

Figure 3 shows the experimental setup (GAME) as well as the computerized tools for the automatic data collection and data analysis:

- The *Facial Expression Analysis Tool* (FEAT [34]) is a connectionist expert system that uses fuzzy rules, acquired from a FACS expert, to automatically measure facial expressions. This expertise is transformed into a network structure by a compiler program. The resulting network is then able to do the classification task, using FACS as the coding language (for more details see [15], [17]). With FEAT we can precisely analyze the dynamics of facial behavior, including intensities and asymmetries.
- The *Interactive Data Elicitation and Analysis* tool (IDEA [41]) provides the possibility to analyze multi media behavioral records of an experimental session and to add new data to the behavioral records. Behavioral records include all data registered during an experimental session, like videotaped data and data automatically registered in an experiment protocol. TRACE [43] can automatically play and analyze a played game in terms of topological measures (e.g., tactical constellations), critical episodes, urgency, etc. The program should eventually approximate the predicted appraisal.

As can be seen in Figure 3, we complement our empirical studies with computer simulation and synthetic modeling:

- The *Facial Action Composing Environment* (FACE [39]) is a tool for creating animated 3D facial expressions in real-time, including head- and eye-movements. The contours of the face are represented with splines, as are the prominent features of the face such as eyebrows and lips, but also wrinkles and furrows (see Figures 1 and 2). The repertoire of facial expressions for the animation is defined on the basis of FACS (for more details see [17], [46]).
- The *Geneva Appraisal Theory Environment* (GATE [38]) is a tool that allows the simulation of different appraisal theories as black box models. This tool enables us to refine or change theoretical propositions incrementally, and it provides immediate feedback on the outcomes. It can also quickly analyze existing data sets for an estimation of the new system's global performance. GATE provides sophisticated facilities to systematically explore such empirical data sets. Furthermore, new empirical data can be obtained directly since the tool incorporates an automatic questionnaire that prompts a subject with a number of questions to determine the appraisal of a situation corresponding to a recalled emotional episode (for more details see [44], [45]).
- The *Autonomous Agent Modeling Environment* (AAME [35]) is a simulation environment for process modeling. It is a tool for designing autonomous agents for use in research and education. The intention was to have an adequate tool that helps to explore psychological and Cognitive Science theories of situated agents, the dynamics of system- environment interactions, and the engineering aspects of autonomous agent design. In the case of the AAME the interesting aspects of such an agent are not so much the modeling of realistic sensors and effectors but the coupling mechanisms between them (for more details see [36], [37]).

4.2 Results Illustrating the Relation between Facial Expression and Appraisal

Within the scope of this article we can only present examples of the current research to illustrate the views on facial expression presented here. More details are published in [16], [17], [18], [19], [45], [46]. The evaluation of situational appraisal profiles allows us to differentiate between different types of identically labeled emotions. We consistently find this result in a) our experimental studies (GAME), in b) studies using the appraisal expert system (GATE) that includes an automatic questionnaire that prompts a subject with a number of questions to determine the appraisal of a situation corresponding to a recalled emotional episode, and in c) judgment studies that make use of synthetic faces and animated facial behavior (FACE). For example, in the experimental computer game interactions, we can consistently distinguish at least three types of anger: a) being angry as an reaction to an unfair event but without blaming anybody, b) being angry and blaming somebody else for having caused the event on purpose, and c) being angry and blaming the other as well as oneself.

Concerning the facial expressions we find a large inter-individual variance in facial expressivity. Furthermore, subjects differ in the variability of the expressions shown. Some subjects show a variety of different facial expressions and situation-specific

repertoires. Additionally, many subjects show "typical" facial expression patterns with a high intra-individual consistency (examples can also be found in [18]). Interestingly, low-expressive subjects do not differ from high-expressive subjects with respect to their reported emotional involvement during the experimental game. Further analyses show that the prototypical patterns described by discrete emotion theorists occur quite rarely and this is also true for subjects who show strong facial reactions. Even more important is the result that the appraisal based approach can better explain and interpret the occurrence of single facial actions. As already mentioned, facial expressions are not exclusively indicators of emotional processes. We also find facial expressions that are signs of cognitive processes (a frown indicating incomprehension), that might or might not be signs of an emotional reaction (anger) at the same time. With an appraisal-based approach we can interpret a frowning, for example, as an indicator of perceiving an obstacle whereas the occurrence of a single facial action is either not interpreted at all by discrete emotion theorists or is interpreted as a possible indicator of a specific emotion, e.g., "possible anger".

Another important advantage of an appraisal based approach is that we can systematically analyze the sources of differences in individual reactions to a specific event in terms of behavior (attack versus avoidance), appraisal (pleasant versus unpleasant, conducive versus obstruct, etc.), and reported feeling (joy, worry, etc.). In addition, we can study individual appraisal tendencies that become evident over different situations and which can be seen as indicators of a more or less stable personal style. For example, some subjects tend to make internal causal attributions (*agency self*) even in situations that are objectively not controllable. Furthermore, we could show how an appraisal-based approach can be applied to understanding and treating emotion pathology and affect disorders [14], [25]. We have formulated concrete predictions concerning disorder specific appraisal biases and deviations from normal emotions including specific facial expression deviations, which can be used to study the underlying – mostly unconscious – processes. Generally, this approach allows us to generate a large number of theoretical predictions that are concrete and testable and that can solidly be based on empirical work.

Results from the studies on facial expression synthesis suggest that subjects perceive the synthetic images and animations generated by FACE in a similar fashion as photographs of real facial expressions [46]. In addition, we used FACE for studying the effect of static versus dynamic presentation of facial expressions. Here, we find that dynamic presentation increases overall recognition accuracy and reduces confusion. One explanation for this encouraging result might be that the facial repertoire was created on the basis of detailed descriptions of the appearance changes produced by each action unit in the FACS manual. As mentioned above, we designed the appearance changes not only for features like eyebrows but also to represent changes in the shape of the facial regions involved and the resulting wrinkles. As a consequence of this strategy, combinations of action units show the same appearance changes in the synthetic face as described in the FACS manual. These changes were not specified but *emerge* from adding the vector information of the respective single Action Units.

5 Conclusion

This article tried to show how an appraisal-based approach might help us to better understand how emotions are expressed and perceived. With respect to human-computer interactions, these two processes refer to the synthesis of “emotional agents” and the analysis of a user’s facial expression.

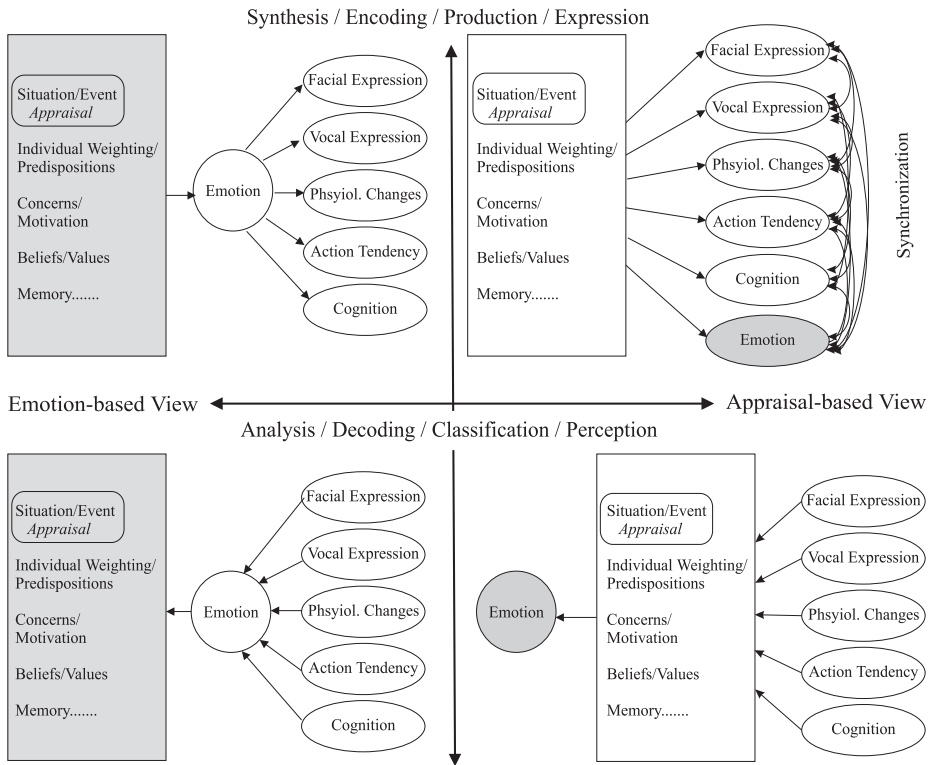


Fig. 4. The two processes of the synthesis and the analysis of facial behavior as studied and conceptualized from an emotion-based and from an appraisal-based viewpoint. Shaded areas indicate parts of the processes that can be omitted within the respective theoretical framework

In our view, componential appraisal theory has some important theoretical, methodological, and ecological advantages over an emotion-based approach. As can be seen in Figure 4, the appraisal-based approach takes not only into account the subjective evaluation of a situation that gives rise to an emotional experience but it directly links the outcome of this appraisal process to the other components of emotions. This allows us to analyze and to synthesize emotional processes on a level of differentiation that goes beyond basic emotions. Since the process of appraisal and reappraisal is not only determined and changed by external situational cues but also by internal cues that reflect the person’s motives, personality, experience etc., those variables should also be considered explicitly.

From the beginning of the early eighties appraisal theories have given important inputs to emotion synthesis and affective user modeling [8]. However, most of these applications only use appraisal theory to implement the “cognitive” component of an emotional interface (for reviews see [23] [24]). The outcome of the appraisal process is then mapped into an emotion category, which determines the synthesis of the respective facial expression pattern. This might be an unnecessary and complicated procedure that also reduces the available information and might bias the analyses. Furthermore, linking facial expression and other components of emotion directly to appraisal dimensions can take advantage of the computational representations that are commonly already defined in the respective applications.

References

1. Ball, G., Breese, J.: Emotion and Personality in a Conversational Character. Workshop on Embedded Conversational Characters. Tahoe City, CA (1998) 83-84 and 119-121
2. Bänninger-Huber, E., Widmer, C.: A New Model of the Elicitation, Phenomenology, and Function of Emotions. In: Frijda, N. H. (ed.): Proceedings of the IXth Conference of the International Society for Research on Emotions. ISRE Publications, Toronto (1996) 251-255
3. Darwin, C.: The Expression of the Emotions in Man and Animals. University of Chicago Press, Chicago (1965) (Original work published 1876 London: Murray)
4. Ekman, P.: Facial Expressions of Emotion: New Findings, New Questions. *Psychological Science* 3 (1992) 34-38
5. Ekman, P., Friesen, W. V.: The Repertoire of Nonverbal Behavior: Categories, Origins, Usage, and Coding. *Semiotica* 1 (1969) 49-98
6. Ekman, P., Friesen, W.V.: The Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto (1978)
7. Ekman, P., Friesen, W. V., Ellsworth, P.: Research Foundations. In: Ekman, P. (ed.): *Emotion in the Human Face*. 2nd edn. Cambridge University Press, New York (1982) 1-143
8. Elliot, C.: I Picked up Catapia and Other Stories: A Multi-modal Approach to Expressivity for 'Emotionally Intelligent' Agents. In: Johnson, W. L. (ed.): *Proceedings of the First International Conference of Autonomous Agents*. ACM Press, New York (1997) 451-457
9. Essa, I., Pentland, A.: Coding, Analysis, Interpretation, and Recognition of Facial Expressions. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19 (1997) 757-763
10. Fridlund, A. J.: *Human Facial Expression: An Evolutionary View*. Academic Press, San Diego (1994)
11. Frijda, N. H.: *The Emotions*. Cambridge University Press, Cambridge New York (1986)

12. Frijda, N. H., Tcherkassof, A.: Facial Expressions as Modes of Action Readiness. In: Russell, J. A., Fernández-Dols, J. M. (eds.): *The Psychology of Facial Expression*. Cambridge University Press, Cambridge (1997) 78-102
13. Izard, C. E.: *The Psychology of Emotions*. Plenum Press, New York (1991)
14. Kaiser, S., Scherer, K. R.: Models of 'Normal' Emotions Applied to Facial and Vocal Expressions in Clinical Disorders. In: Flack, Jr., W. F., Laird, J. D. (eds.): *Emotions in Psychopathology*. Oxford University Press, New York (1998) 81-98
15. Kaiser, S., Wehrle, T.: Automated Coding of Facial Behavior in Human-Computer Interactions with FACS. *Journal of Nonverbal Behavior* 16 (1992) 67-83
16. Kaiser, S., Wehrle, T.: Situated Emotional Problem Solving in Interactive Computer Games. In: Frijda, N. H. (ed.): *Proceedings of the IXth Conference of the International Society for Research on Emotions, ISRE'96*. ISRE Publications, Toronto (1996) 276-280
17. Kaiser, S., Wehrle, T.: Facial Expressions as Indicators of Appraisal Processes. In: Scherer, K. R., Schorr, A. (eds.): *Appraisal Theories of Emotions: Theories, Methods, Research*. Oxford University Press, New York (In press)
18. Kaiser, S., Wehrle, T., Edwards P.: Multi-Modal Emotion Measurement in an Interactive Computer Game: A Pilot-Study. In: Frijda, N. H. (ed.): *Proceedings of the VIIIth Conference of the International Society for Research on Emotions. ISRE Publications, Storrs* (1994) 275-279
19. Kaiser, S., Wehrle, T., Schmidt, S.: Emotional Episodes, Facial Expression, and Reported Feelings in Human-Computer Interactions. In: Fischer, A. H. (ed.): *Proceedings of the Xth Conference of the International Society for Research on Emotions, ISRE'98*. ISRE Publications, Würzburg (1998) 82-86
20. Lazarus, R. S.: *Psychological Stress and the Coping Process*. McGraw Hill, New York (1966)
21. Leventhal, H., Scherer, K. R.: The Relationship of Emotion to Cognition: A Functional Approach to a Semantic Controversy. *Cognition and Emotion* 1 (1987) 3-28
22. Lien, J.J., Kanade, T.K., Zlochow, A.Z., Cohn, J.F., Li, C.C.: A Multi-Method Approach for Discriminating Between Similar Facial Expressions, Including Expression Intensity Estimation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Santa Barbara, CA* (1998)
23. Pfeifer, R.: Artificial Intelligence Models of Emotion. In: Hamilton, V., Bower, G. H., Frijda, N. H. (eds.): *Cognitive Perspectives on Emotion and Motivation*. Kluwer Academic Publishers, Dordrecht (1988) 287-320
24. Picard, R. W.: *Affective Computing*. The MIT Press, Cambridge (1997)
25. Roseman, I. J., Kaiser, S.: Applications of Appraisal Theory to Understanding and Treating Emotion Pathology. In: Scherer, K. R., Schorr, A. (eds.): *Appraisal Theories of Emotions: Theories, Methods, Research*. Oxford University Press, New York (In press)
26. Roseman, I. J., Wiest, C., Swartz, T. S.: Phenomenology, Behaviors, and Goals Differentiate Discrete Emotions. *Journal of Personality and Social Psychology* 67 (1994) 206-221

27. Russell, J. A., Fernández-Dols, J. M. (eds.): *The Psychology of Facial Expression*. Cambridge University Press, Cambridge (1997)
28. Scherer, K. R.: On the Nature and Function of Emotion: A Component Process Approach. In: Scherer, K. R., Ekman, P.(eds.): *Approaches to Emotion*. Lawrence Erlbaum, Hillsdale (1984) 293-318
29. Scherer, K. R.: What Does Facial Expression Express? In: K. Strongman (ed.): *International Review of Studies on Emotion*, Vol. 2. Wiley, Chichester (1992) 139-165
30. Scherer, K. R.: Studying the Emotion-Antecedent Appraisal Process: An Expert System Approach. *Cognition and Emotion* 7 (1993) 325-355
31. Smith, C.A., Ellsworth, P.C.: Patterns of Cognitive Appraisal in Emotion. *Journal of Personality and Social Psychology* 48 (1985) 813-838
32. Smith, C. A., Scott, H. S.. A Componential Approach to the Meaning of Facial Expression. In: Russell, J. A. Fernández-Dols, J. M. (eds.): *The Psychology of Facial Expression*. Cambridge University Press, Cambridge (1997) 229-254)
33. Takeuschi, A., Naito, T.: Situated Facial Displays: Towards Social Interaction. In: CHI'95, *Human Factors in Computing Systems* (1995) 450-454
34. Wehrle, T.: *The Facial Expression Analysis Tool (FEAT)* Unpublished Computer Software. University of Geneva, Switzerland (1992/1996)
35. Wehrle, T.: *The Autonomous Agent Modeling Environment (AAME)* Unpublished Computer Software. University of Geneva, Switzerland (1993)
36. Wehrle, T.: *Eine Methode zur Psychologischen Modellierung und Simulation von Autonomen Agenten*. Unveröff. Dissertation an der Philosophischen Fakultät I, Universität Zürich (1994a)
37. Wehrle, T.: New Fungus Eater Experiments. In: Gaussier, P., Nicoud, J.-D. (eds.): *From Perception to Action*. IEEE Computer Society Press, Los Alamitos (1994b)
38. Wehrle, T.: *The Geneva Appraisal Theory Environment (GATE)*. Unpublished Computer Software. University of Geneva, Switzerland (1995a)
39. Wehrle, T.: *The Facial Action Composing Environment (FACE)*. Unpublished Computer Software. University of Geneva, Switzerland (1995b/1999)
40. Wehrle, T.. *The Geneva Appraisal Manipulation Environment (GAME)*. Unpublished Computer Software. University of Geneva, Switzerland (1996a)
41. Wehrle, T.: *The Interactive Data Elicitation and Analysis (IDEA)*. Unpublished Computer Software. University of Geneva, Switzerland (1996b)
42. Wehrle, T.: Motivations behind modeling emotional agents: Whose emotion does your robot have? Paper presented at the Workshop on Grounding Emotions in Adaptive Systems, hold at the 5th International Conference of the Society for Adaptive Behavior (SAB'98), University of Zurich, Switzerland (1998)
43. Wehrle, T.: *Topological Reconstruction and Computational Evaluation of Situations (TRACE)*. Unpublished Computer Software. University of Geneva, Switzerland (1999)
44. Wehrle, T., Scherer, K. R.: Potential Pitfalls in Computational Modelling of Appraisal Processes: A Reply to Chwelos and Oatley. *Cognition and Emotion* 9 (1995) 599-616

45. Wehrle, T., & Scherer, K. R.: Towards Computational Modeling of Appraisal Theories. In: Scherer, K. R., Schorr, A. (eds.): *Appraisal Theories of Emotions: Theories, Methods, Research*. Oxford University Press, New York (In press)
46. Wehrle, T., Kaiser, S., Schmidt, S., Scherer, K. R. Studying the dynamics of emotional expression using synthesized facial muscle movements. *Journal of Personality and Social Psychology* 78 (2000)

A Cognitive Approach to Affective User Modeling

Carlos Martinho¹, Isabel Machado², and Ana Paiva¹

¹ IST-Technical University of Lisbon and INESC

² CBLU - University of Leeds and INESC

Rua Alves Redol, 9, 1000 Lisboa, Portugal

{carlos.martinho,isabel.machado,ana.paiva}@inesc.pt

Abstract. This paper proposes a framework for cognitive-based affective user modeling. Such framework relies on the idea that, to model affect in user models, one can use the situations experienced by the user as well as the observable behaviour of the user. These situations and behaviours, when interpreted under the light of a cognitive theory of emotions, will allow us to make inferences on the possible affective states of the user. Such type of approach will be here called cognitive-based affective user modeling. To describe the framework we will use a cognitive theory of emotions (the Ortony, Clore and Collins theory of emotions) as the basis for the representation of the affective states of the user. We will also illustrate the use of this modelling in the context of a virtual theatre, where children engage in collaborative story creation.

1 Introduction

User Modeling (UM) is an essential technique to attain personalised communication between Humans and interactive computer systems. Over the years, the question of adaptation to the user has been approached with a focus on acquiring and representing the user's preferences, goals, and state of knowledge. Such elements compose a user model and are normally inferred through the use of diverse types of approaches, such as machine learning techniques to capture users' preferences, or cognitive diagnosis techniques to capture users' misconceptions in a domain knowledge [19].

More recently, and influenced by the emerging field of Affective Computing, i.e. "computing that relates to, arises from or deliberately influences emotions" [16], a new area called **Affective User Modeling** (AUM) has appeared and is now giving its first steps. In [4], AUM is defined as the capability for a system to model the user's affective states. Indeed, as Humans interact directly with computers, it is critical for this interaction to be empowered with affective components, not only in the expression of affect by the system but also in the capture of and reaction to the affective states of the user.

A substantial amount of work has already been done in the acquisition of a user's affective states (e.g. Affective Computing group at the MIT Media Lab [17]). Most of the research development is based on multi-modal forms

of input as affective *wearables* [16], speech recognition [1] and facial expression recognition [20]. But if the physiological aspects of emotions can be identified and collected through the use of different types of sensors, the typical interactions established nowadays with computers applications still use traditional means, such as keyboard, joystick and mouse input.

However, while interacting with computer applications, users communicate their affective states not only through body expressions and physiological signals, but also through their behaviour. In a well defined context (e.g. a computer learning situation), the users' behavioural responses may be a path to predict, recognise and interpret the user's affective states. Such approach, here named **Cognitive-Based Affective User Modeling** (CB-AUM), has already been proposed by Elliott *et al.* [4].

CB-AUM should not only be regarded as a way of ascribing emotional states to the users, but also as a different component in the user modeling process which, when combined with affective *wearables*, speech recognition, and other multi-modal components of user input, will make AUM much more powerful. Indeed, as argued in [16], the best recognition is likely to come from the combination of the different modalities and including not only low-level signal recognition but also higher-level reasoning about the situation.

This paper is organised as follows. First, we will present a scenario for the exploration of the approach and provide with the motivation behind this work. Then, we will describe what we mean by cognitive-based affective user modeling by presenting a framework for it, based on the cognitive theory of emotions of Ortony, Clore and Collins (OCC), and positioning it within the broader field of user modeling. Afterwards, we will illustrate how such an approach can be used within the context of a virtual theatre, where children control their own characters in a collaborative story-creation environment. Finally, we will make some concluding remarks and draw some plans for future work.

2 Scenario and Motivation

To better explain the framework for AUM, we will first present a scenario of a typical child interaction with a collaborative virtual environment (CVE) for story creation: *Teatrix*¹ [18]. *Teatrix* aims at providing effective support for young children (7-9 years old) developing their notions of narrative, through the dramatisation of several situations.

Through *Teatrix*, each child collaborates with other children to create a story. Each one controls a character by selecting actions from a list of available actions for her/his character to perform in the 3D story world (see Figure 1). Additionally, each child can also induce some emotional state changes in her/his character through a tool called "hot-seating" (see [11]).

¹ Teatrix is one application developed within the NIMIS project, a EU funded project within the I3-ESE, contract n.29301.



Fig. 1. *Teatrix* Environment

A fundamental idea in *Teatrix* is that when a child chooses a character to direct, s/he is also implicitly choosing a role to play (e.g. the *Hero* or the *Villain*) with a predefined set of goals and behaviours (such as: “defeat the villain”).

In *Teatrix*, there is no well defined learning progression that would allow us to assess the user achievements. Thus, the user modelling component does not contain the usual domain knowledge elements found in many user modelling systems. However, since the story creation process is mediated by a narrative guidance system, some information about the degree of control of the character, about the motivation for the story and, most importantly, about the involvement a child has in the story, are elements that will help the personalisation of the guidance system.

So, for instance, when a child is controlling a *Hero* character, s/he will perform actions aiming at defeating the *Villain*. David Wood and Janet Grant in their book *Theatre for Children* emphasize the fact that children develop empathic relations with the characters of a play [21]. Children enjoy taking sides, identifying themselves with a good character or being mad with the behavior of a untruthful character. They are prone to get emotionally involved with the characters. Taking these facts into account, one can argue that not only the characters are acting out in the story but also that the children themselves become engaged in the enactment activity. In other words, in *Teatrix*, we can gather the information about the child’s affective state not only by looking at their degree of control over the characters, but also by considering their characters goals,

emotional state and reactions. With this information, we plan to provide them with a more motivational experience and a richer narrative support.

3 A Starting Point: User Modelling Research

Since the seventies, both the learner and user modeling research communities have dwelt into the problem of acquiring, representing and using users' characteristics and attitudes to personalise the interaction. From a focus on how to infer students' states of knowledge and diagnose their misconceptions, to how to obtain the users' preferences, many systems have been build and used (see [10] for a good collection, or more recently [5] and [8]). Moreover, and to deal with different types of demands, UM shells were built as a way to facilitate applications to integrate UM functionality (for example and among others: BGP-MS [9], UM-toolkit [7], TAGUS [15]). Most of these UM systems and shells allow the user model to contain different types of information such as: knowledge, beliefs, characteristics, and preferences (see [7]).

In general, the different information types found in traditional user models are:

Characteristics: normally captured within a profile of the user, such as age, gender, and interests. This information can be relevant if stereotypes are used to infer other more specific information, such as preferences.

Domain Knowledge: the beliefs of the user about a domain of knowledge. This type of content is relevant for Intelligent Learning Environments and aims at responding to the student's state of knowledge and providing with adequate advice.

Preferences and Needs: normally used in interactive systems to advice the user, or in interface agents to select the most appropriate information to provide. Most of the information retrieval support agents are based on the representation of the information "needs" or "preferences" of the users (among others [3], [2]).

Goals and Plans: used in problem solving or task execution situations, both on learning environments and help systems. The need for goals and plans is an immediate result of the need to support the users to achieve *adequately* their tasks. A large and important amount of work has been done in the area of acquiring the users' plans (e.g. [6]).

Most of these elements can be obtained, with some degree of certainty, by the analysis of the user's behaviour. To express this uncertainty of the models, numerical factors are often used.

4 A Framework for Cognitive-Based Affective User Modeling

In AUM, we want to represent, acquire, and hypothesize about the user's emotional states. Hence, we need: (1) to include a representation of the emotions or

affective states in the user model; and (2) to develop techniques to make predictions and inferences around these emotions. So, the main questions to address are:

1. How to acquire the user's affective states?
2. How to represent them?
3. How to use these representations effectively?

In the next subsections, we will try to provide with answers to these questions.

4.1 Representing Affective States

We will be ascribing emotions to the user based on the user's behaviour and on the events of the world. Hence, we need a cognitive theory of emotions that consider and work with such stimuli. From the vast array of theories of emotions stating how emotions are generated, we chose the OCC theory [14] as the support for the construction of our affective user model. In the remaining of this section, we will describe how the OCC theory can be used for the representation of the user's affective states.

Based on the OCC theory, the user affective model can be characterised by two type of data: the user *emotional profile*, and the *emotions* themselves.

The following subsections explain each one of them.

a) User Emotional Profile. The emotional profile states the assumed emotional conditioning of the user and is constituted by:

Emotional class thresholds: representing the assumed emotional “resistance” towards the different classes of emotions (Figure 2 represents the OCC structure of the emotion classes).

As an example, emotion thresholds can model how easily the child can be disappointed during story creation. Character's goal failure can give cues on this threshold value when a child persists on guiding her/his character to achieve a particular goal through a set of actions but never does achieve the intended goal. The amount of effort produced by the child can be used to state the level of resistance of the child towards disappointment.

Emotional class decays: represent how long the emotions assumed to be experienced by the user last.

Although these values are very difficult to predict during the interaction, they are strongly related with the level of empathy that a child developed for the story characters. This empathy can be inferred through various means as the regularity of character/role choice, or direct questionnaires.

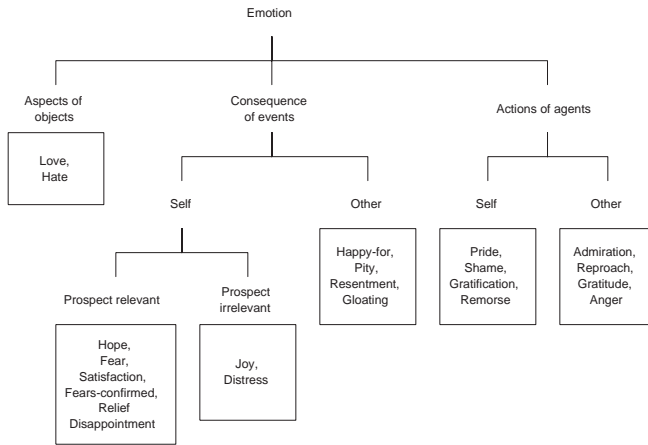


Fig. 2. OCC Classes of Emotion

b) User Emotions. Emotions stored by the AUM system are of two types:

Potential emotions: represent which particular classes of emotions the current situation is likely to provoke according to the child's inferred attitudes, goals, and standards of behaviors. Additionally, they state how strong those potential emotions are (i.e. their *potential*).

For instance, the child will be considered to feel reproach if a particular character that was cooperating with her/his character (i.e. both characters had common goals), suddenly changes its behaviour and enters directly in conflict with her/his character (i.e. the characters have now conflicting goals).

Active emotions: whenever the potential of an emotion overcomes the assumed resistance of the user towards this class of emotion, the emotion is said to be *active*, and the user is assumed to be experiencing it. The user model stores these emotions, along with their *intensity*, that is, the distance between the user's assumed threshold and the potential of the respective potential emotions. The greater this distance, the higher the intensity of the experienced emotion will be assumed to be.

Besides potential and intensity, each emotion is characterised by two other dimensions: the *cause*, that is, the event(s) which provoked it, and the *direction*, that is, towards who or what the emotion is directed.

4.2 Appraisal Structure

To activate the assumed emotions of the user with an accurate intensity, the UM system must consider the mechanism by which emotion-inducing stimuli are

appraised. In the OCC theory, this evaluation is based on a set of representation elements — the *appraisal structure* — which provide information on how to build a valenced interpretation of a situation from the user point of view.

The appraisal structure is an assumption on the way people perceive the world, and how this perception leads to the activation of emotions. Central to the work of OCC is the notion that emotions are valenced reactions which result from three different perspectives of seeing the world: the events happening in the world; the user and other agents' actions in the world and; the objects that exist in the world.

Transcribing the theory to the UM environment is almost direct. Events are appraised according to their pleasingness and desirability with respect to the user and other agents' assumed goals, and may activate *consequence of events* emotions such as hope and disappointment. Actions are appraised with respect to the user's assumed standards of behavior, and may activate *action of agents* emotions such as shame and pride, according to the approval or disapproval of these same actions. Finally, objects are appraised based on how the user is assumed to like them, launching *aspect of objects* emotions such as love and hate. All the classes of emotion represented in Figure 2 fall under one of those three perspectives.

The following subsections present how this appraisal structure can be stored and used by the UM system, in the case of the presented scenario.

World Events and User Goals. All events must be appraised in the light of the user's goals, to infer potential emotions related with consequences of events. If the OCC theory clearly states when to launch the different emotions when faced with desirable or undesirable events, how can the AUM system classify an event as desirable or undesirable?

Assuming that the relative importance of success and failure (which defines the goal desirability) of user assumed and system goals are well defined, and that goals are organised as a inter-relation network (e.g., as the goal graph used in [12]) stating how each goal realisation affects the realisation of the other goals of the network, the appraisal process should follow the following two simple guidelines:

- Events that lead to or facilitate the realisation of desirable goals, and events that prevent or inhibit the achievement of undesirable goals are desirable.
- Events that lead to or facilitate the realisation of undesirable goals, and events that prevent or inhibit the achievement of desirable goals are undesirable.

In both cases, the degree of (un)desirability is proportional to the goal importance and the degree to which the event contributes to (or prevents) the goal achievement. Those values are, however, rarely always clearly perceived by the user, and her emotions will be generally experienced with different intensities than if the user could “see the all picture”.

In a story creation process, it is quite difficult for the child to evaluate how a single directive will contribute for the story progression. Using the example of a child controlling a *Boy* character fighting against the evil *Witch*, which actions aim at proving himself worthy of the *Fairy*'s help. S/he does not know what kind of powers the *Fairy* will empower the character with. Hence, s/he does not either know how much realising the goal "obtain *Fairy*'s help" contributes for her/his primary goal "defeat the *Witch*". Therefore, when designing the network, we must use goal values reflecting the child assumed partial view rather than the narrative guidance system overall evaluation of each situation. It is important to note that, however, as the children use the system through time, this "surprise effect" will tend to disappear, as the children learns the different possibilities of the system... Thus, ideally, the system should take into account the experience of the child with the environment. This is clearly a very sensitive point of development.

Nevertheless, the OCC theory considers other more easily definable dimensions affecting the emotional appraisal as, for instance, the effort invested in achieving the goal. Those variables may compensate the mentioned lack of information.

Another important factor must be considered. In UM systems, the usual interpretation of a user goal is "a motivation to act". However, users do not only have action-directed goals: they also have interest goals, which are an important piece in understanding the user's behaviour.

As an example, children have preferences in terms of what roles they most like to perform when creating a story. This information can enhance the system with features that would provide the children with a more engaging experience at each new interaction. For instance, the system can propose to the child the characters/roles matching its assumed preferences, using as a basis her/his selected actions and character emotional changes in past stories: a child who keeps on trying to make other characters eat poisoned apples will probably like to play *Villains* more than *Heroes*.

If we intend to use currently available UM techniques that capture the user goals, we will have to consider an extension to this "usual" notion of the goals to capture also interest goals.

However, this automatically raises a problem: if interest goals are not always directly the cause for an action, how can a system ascribe these higher level goals to the user? A possibility would be to use *direct questioning* or *self reflection*, grounding it completely on an empirical basis or aiming at the definition of the user personality through personality trait theory forms [13].

User Attitudes and Standards. As mentioned above, the appraisal structure also contains information on the user’s assumed attitudes² towards objects, as well as her assumed standards of behaviour.

Attitudes towards world objects can be extrapolated from the user behaviour. For instance, the system may rely on observation and previously gathered data (e.g. direct questioning) to infer that the user have preferences upon the type and properties of the presented props, and assume that the user is experiencing love and hate emotions when using these materials.

There is already a large amount of work done in the UM community focusing on the ascription of preferences to the user, which can be seen as the attitudes of the user towards some entities of the world, stating a measure of the appeal of these world objects for the user. Hence, we can rely on such techniques to obtain preferences as a form to acquire the user attitudes.

Standards are a measure of how “good” or “bad” an action of the user or system is, based on, among others, the user assumed moral, justice values, and ethical code of conduct. Similarly to interest goals, standards of behaviour may be difficult to obtain from user behaviour observation alone. If, from the teaching point of view, right and wrong procedures are well defined, the same can not be said about user standards. Standards will be mostly assumed from the cultural background of the target users of the system. The system may, however, and as always, use any of the techniques mentioned earlier to test the user assumed standards.

4.3 Two-Stage Affective Cycle

The CB-AUM system must be able to simulate the user’s appraisal process. However, it does not yield much practical use unless there are some reasoning and inference mechanisms that provide with the power of drawing inferences based on the maintained affective model. Furthermore, the system must also have the capacity of updating the model according to the comparison of system formed hypotheses about user behaviour, and the actual observation of her performance.

Thus, our framework can be viewed as a two-stage process pipeline — the *Affective Cycle* — represented in Figure 3, that should be interpreted as follows. If there are any expectation on the user behaviour, based on an assumed hypothetical previous change in her affective state, the current observed behaviour is compared to it. According to the confirmation or not of the predicted behaviour, changes are performed in the affective user model, to reinforce or inhibit certain parameters of the affective user model. The system then performs the second stage of the affective cycle: the simulation of the appraisal. The CB-AUM system infers, based on current observed world events (system events and user behaviour actions) as well as her current affective model (mainly, her appraisal structure),

² In this paper, “attitude” is used in the sense given by Ortony, Clore and Collins in [14]: “(...) attitudes, which we view as the **dispositional** liking (or disliking) one has for certain objects, or attributes of objects, without reference to standards or goals.” - p46.

what affective changes will the observed events activate, and what actions are expected to occur based on these affective changes. These expectation will then be used by the reinforcement stage in the next affective cycle.

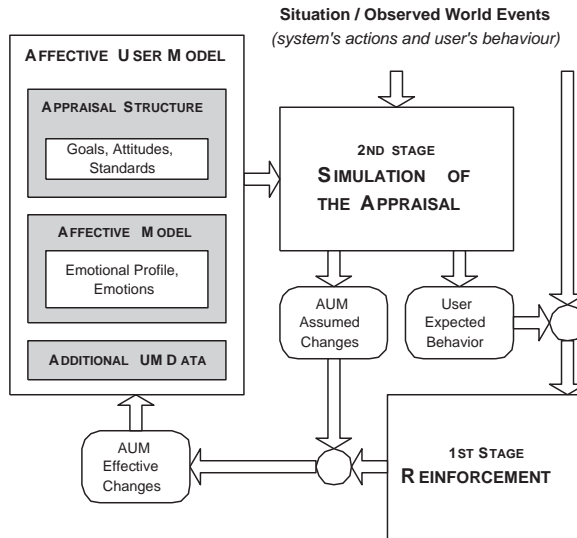


Fig. 3. 2-Stage Affective Cycle

The next subsections discuss the two stages.

Simulation of the Appraisal. This stage is mainly interested on how to use the user's appraisal structure to ascribe an emotion to a user, and therefore be able to maintain a prediction on the user emotional state.

As previously described, the appraisal structure characterises the typical reactions of a user to world situations according to three dimensions: goals, attitudes and standards. These reaction will trigger different emotions according to OCC theory, that the system assumes to be experienced by the user. However, how does the system evaluate each world event?

Our appraisal stage is, itself, also a two-step process. The first step is to identify the relevance of the observed events for the user. All events that are perceived by the user are cross-checked with the user's appraisal structure.

For instance, characters and props that the child likes (stored in her attitudes), and events or characters' behaviours that the child does not like (stored in her standards structure), are filtered from all the perceived events. All other elements are dumped.

Once the relevant events are filtered out, the second stage of the appraisal is to evaluate them in the light of the OCC theory of emotions, and to evaluate each emotional parameter. Implementation-wise, this stage relies heavily on pattern matching techniques. An example of such implementation can be found in [12].

Reinforcement of the Hypotheses. Implementation-wise, this phase relies heavily on reinforcement learning techniques to maintain the coherence of the user's emotional profile. This stage should be combined with user attuned affective *wearables* and other types of affective input to achieve a more reliable model of the affective states of the user.

5 Concluding Remarks

In this paper, we presented a first study on how to build affective user modeling systems through the inclusion of cognitive-based techniques based on a cognitive theory of emotions. Using the Ortony, Clore and Collins cognitive theory of emotions, we described a framework based on a two stages processing and discussed the necessary type of information such systems must keep, as well as some of the initial problems we must face when prototyping them.

We further pointed how this cognitive-based approach, rather than replacing physiological signal processing, can be combined with such technology to produce more robust systems. The cognitive part would provide with the tools to predict affect based on world interpretation, and affective input would provide with elements to reinforce or inhibit those predictions based on confirmed physiological signals, in addition to the observation of the user's behaviour.

References

1. G. Ball and J. Breese. Modelling the emotional state of computer users. In Fiorella di Rosis, editor, *Workshop on Personality and Emotion in User Modelling - UM'99*. UM 99, 1999. 65
2. N. Belkin. Intelligent information retrieval: whose intelligence? Technical report, Rutgers University, 1998. 67
3. D. Billsus and M. Pazzani. A hybrid user model for news story classification. In Judy Kay, editor, *Proceedings of User Modeling Conference UM'99*. Springer, 1999. 67
4. C. Elliott, J. Lester, and J. Rickel. Lifelike pedagogical agents and affective computing: an exploratory synthesis. In Mike Wooldridge and Manuela Veloso, editors, *AI Today*. Springer - Lecture Notes in Artificial Intelligence, 1999. 64, 65
5. A. Jameson and C. Paris. *Proceedings of User Modeling Conference UM'97*. Springer, 1997. 67
6. H. Kautz. A circumscriptive theory of plan recognition. In Philip Cohen, Jerry Morgan, and Martha Pollack, editors, *Intentions in Communication*. MIT Press, 1990. 67
7. J. Kay. The um toolkit for reusable, long term user models. *User Modeling and User-Adapted Interaction*, 4(2), 1995. 67

8. J. Kay. *Proceedings of User Modeling Conference UM'99*. Springer, 1999. 67
9. A. Kobsa and W. Pohl. The user modeling shell system bgp-ms. *User Modeling and User-Adapted Interaction*, 4(2):59–106, 1994/1995. 67
10. A. Kobsa and W. Wahlster. *User Models in Dialogue Systems*. Springer-Verlag, 1989. 67
11. I. Machado and A. Paiva. The child behind the character. In Kerstin Dautenhahn, editor, *(to appear in) AAAI Fall Symposium on Socially Intelligent Agents*. AAAI Press, 2000. 65
12. C. Martinho. Emotions in motion. Master's thesis, Instituto Superior Tecnico, Universidade Tecnica de Lisboa, 1998. 70, 74
13. W. Mischel. *Introduction to Personality*. Harcourt Trade Publishers, 6th edition, October 1998. 71
14. A. Ortony, G. Clore, and A. Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, 1988. 68, 72
15. A. Paiva and J. Self. Tagus- a user and learner modeling workbench. *User Modeling and User-Adapted Interaction*, 4(2), 1995. 67
16. R. Picard. *Affective Computing*. The MIT Press, 1997. 64, 65
17. R. Picard. Affective computing research group. Web page, MIT Media Lab, <http://www.media.mit.edu/affect/>, 1999. 64
18. R. Prada, I. Machado, and A. Paiva. Teatrix: Virtual environment for story creation. In *Intelligent Tutoring Systems, ITS'2000*. Springer, 2000. 65
19. J. Self. Model-based cognitive diagnosis. *User Modeling and User-Adapted Interaction*, 3:429–447, 1993. 64
20. T. Wehrle and S. Kaiser. Emotion and facial expression. In A. Paiva, editor, *Affective Interactions (this volume)*. Springer, 2000. 65
21. D. Wood and J. Grant. *Theatre for Children - A Guide to Writing, Adapting, Directing and Acting*. Faber and Faber, 1997. 66

Affective Appraisal *versus* Cognitive Evaluation

in Social Emotions and Interactions

Cristiano Castelfranchi

National Research Council - Institute of Psychology
Division of "Artificial Intelligence, Cognitive and Interaction Modelling"
castel@ip.rm.cnr.it

Abstract I claim that two important aspects of emotions are usually missed by current computational models and uses. On the one side, human emotions are *complex and rich mental states*, not simple reactive mechanisms. They have rich *cognitive ingredients*, in particular “evaluations”. I will propose some brief example of such a necessary “cognitive anatomy” (in terms of beliefs and goals) of complex social emotions (for ex. shame) and of the “meaning” of their expression. On the other side, emotions are *felt*; and we do not have a model about what it does mean to “feel” something, about the unavoidable role of a “body” in this, and about the necessary function of the feeling component in the emotional process. In this perspective, in particular it becomes very important the distinction between a true “cognitive evaluation” and merely *intuitive and implicit affective appraisal*. I exemplify this distinction in the analysis of “trust”. Finally, I wonder about the importance of the affective appraisal Vs the explicit evaluations in Affective Computing and in HC and computer mediated interactions. Are these cognitive complex emotions, are these felt and embodied emotions, are explicit evaluations or intuitive appraisals useful and why?

I also discuss a bit the relationships between emotion and motivation, and between emotion and personality, that on my view are currently quite mixed up.

1. Two Missed Aspects of Computing Emotions

Something important is usually missed by current computational and robotic models and uses of emotions.

- On the one side, human emotions are *complex and rich mental states*, not simple reactive mechanisms. They have rich *cognitive ingredients*.

I will propose some examples of such a necessary “cognitive anatomy” (in terms of beliefs and goals) of complex social emotions, and of the “meaning” of their expression as communication about the mental state.

- On the other side, emotions are *felt*; and we do not have a model of what it does mean to “feel” something, about the unavoidable role of a “body” in this, and about the necessary function of the feeling component in the emotional process.

I will illustrate these two problems,

- by considering in some emotions and affects (envy, shame, and trust) the relevance of their cognitive anatomy;
- by illustrating the distinction -crucial for a theory of emotions and of affective reactions- between intuitive and implicit appraisal and cognitive evaluation;
- by providing some hints about what it should mean/imply “to feel” something.

In this structural and functional analysis of emotions I will also discuss a bit the relationships between emotion and motivation, and between emotion and personality (see Appendix), that in my view are currently quite mixed up.

In the last part I will reason about the questionable importance of the cognitive stuff, of the affective appraisal Vs the explicit evaluations, and of the felt emotions, in Affective Computing and in HC and computer mediated interactions.

2. Cognitive Aspects in Emotion: Emotions as Complex and Rich Mental States, not Simply as Reactive and Expressiv Devices¹

Emotion forms a complex, hybrid subjective state (*state of mind*). This state is described as “hybrid” in that the emotions are constituted by the integration of mental, somatic, cognitive and motivational components. A constituent element of each complex emotion is the “mental state”, an integrated *structure* of assumptions and goals which is closely linked to its functional nature, determines its intensional nature and classification, and accounts for its activation and motivation.

Intension

Many emotions have *intension*, i.e. they are about something. They usually have an *object* (something which arouses emotion) and often a *target*, i.e. someone they are directed towards. For example, we envy someone for something, get angry with someone over something, feel ashamed or guilty towards someone about something.

Value or Valence

The emotions can also be divided for the most part into “positive” and “negative”. These terms are, however, ambiguous in that they confuse different aspects. Positive and negative emotions can be understood on the one hand as those which are subjectively pleasant or unpleasant, and on the other as emotions regarding success in achieving goals (joy, pride) or failure (fear, shame, guilt). Emotions can also be regarded as positive or negative in that they are culturally praised and encouraged or criticized and discouraged (e.g. envy), or in the sense that the individual avoids them (e.g. tenderness, because it makes him “vulnerable”) or pursues them (e.g. the fear involved in taking risks, because it makes him feel “alive”).

In any case, the general thesis is that *the emotions indicating failure (actual or threatened) to achieve a goal are subjectively negative, i.e. unpleasant*. It is, however, essential to have a clear idea of the goals in question (see later, and [25]).

¹This part of the paper is an abridged translation of a chapter of an Italian book: *Maria Miceli e Cristiano Castelfranchi "Le difese della mente"* (Defences of Mind), NIS, Roma, 1995. I would like to thank Maria for her precious contribution to both the content and the form of this chapter.

The Constituent Elements of the Emotions

The basic constituent elements of the emotions are beliefs, evaluations, goals, arousal - i.e. somatic activation and its *proprioception* - and the “tendency towards action” or conative component.

2.1 Beliefs

Beliefs play three fundamental roles in the emotions.

i. Activating beliefs. Many emotions are activated not by external stimuli but by representations, which can also be endogenously generated (e.g. through inference). These representations can be either of the perceptive-motor type (e.g. images) or propositional (e.g. “John is dead”, “I might be found out”). Such beliefs activate and partially account for emotion in that they are consistently connected with the individual’s conduct. In the case of guilt, for example, the subject entertains a complex structure of interconnected beliefs: that something has happened; that this injures someone; that this injury is not “deserved” by the person suffering it; that the subject’s own actions or negligence are responsible for the injury; that it could have been avoided. This *configuration* of beliefs produces the sense of guilt and is an integral part thereof. They form a lasting part of the “mental state” accompanying the sense of guilt, are involved in evaluation, causal attribution and categorization (as we shall see), and are connected with the goals activated.

In many emotions, if not in all, some of the activating beliefs are very special and clarify one of the essential links between emotions and goals. Emotive reactions appear to be connected with the monitoring of certain goals, being triggered when these goals are compromised/threatened or achieved/announced. *Failure beliefs* therefore unquestionably form part of activating beliefs.

Positive or negative *evaluations* (see 4.1) very often form part of activating beliefs. For example, the sense of guilt involves a negative evaluation of the event (injury) and hence of one’s action and of oneself; the same is true for shame relatively to some feature or action of the subject not necessarily involving any responsibility.

There can be no doubt that failure/achievement beliefs are themselves implicitly or explicitly evaluative beliefs.

ii. Causal attribution beliefs. Activating beliefs produce a somatic reaction, a more or less specific *arousal* of each emotion. This reaction is an essential component of “how one feels”, of the subjective aspect of the emotion, but does not merely accompany the ideas of images that activate it. In a cognitive agent, a far more explicit nexus is established. The mind *interprets* these stimuli and attributes them precisely to the emotive state. It assumes that the changes and sensations involved are due to a given emotion and to the activating beliefs and images belonging to it. In other words, it causally attributes the arousal and the sensations to the emotion (and in particular to the activating stimuli and representations).

We shall therefore use the term “attributional beliefs” in this context to refer to the attribution of a somatic reaction to a particular emotive state (not to activating beliefs of the causal type, e.g. beliefs regarding the attribution of responsibility in the case of guilt).

iii. Categorization beliefs. Categorization beliefs are no less important and typical of the emotions in cognitive agents. The subject interprets, recognizes and labels his state of mind as a certain emotion (one which his language and personal vocabulary enable him to express), saying to himself “I’m furious with John”, “I’m sad”, “I love Lucy” or “What a disgrace”. These categorization beliefs (which are,

in a certain sense, meta-level beliefs) are closely connected with the other beliefs. Categorization is only possible in the presence of a specific configuration of beliefs and attribution of the associated sensations. In other words, a given interpretation of facts and sensations is required if we are to feel a given emotion, which for human beings means “recognizing” that emotion.

In defending oneself against certain emotions or against the suffering involved [36], one attacks precisely some of these beliefs, thereby dismantling the assumptions, ingredients or recognition-awareness of the emotion. It is, of course, also possible to work on one’s goals (more precisely, on the beliefs supporting them), which are again often constituent elements of the emotion in question.

2.2 Goals

The relationship between emotions and goals is also structured and complex. The emotions themselves *are translated* into goals. They have the role of *monitoring* the pursuit and achievement of goals. Finally, they *activate* goals.

i. *Emotions as goals.* The individual can perform or avoid performing an action *in order to* feel or not to feel a given emotion. He can, for example, give a present in order to feel the pleasure caused by the satisfaction or gratitude of the recipient, or forbear from doing injury in order to avoid feeling guilt. In behaviourist terms, the (positive or negative) emotions are therefore often a (positive or negative) reinforcement for the adoption or avoidance of given behaviour. For this reason they play an important role in learning. A “successful” (or ineffective or harmful) action or plan will subsequently be pursued (or avoided) not only in the light of memories, inferences and evaluations regarding outcome (success or failure), costs of pursuit and side effects, but also (and in more direct fashion, without the mediation of reasoning) in order to experience (or avoid) the emotions associated with the outcome of the behaviour in question.

ii. *The emotions monitor goals.* In addition to the possibility of themselves becoming goals, the emotions perform the role of *informing* the individual as to whether his goals are compromised or achieved (threatened or “promised”). Fear, anxiety, shame, guilt, surprise, joy, pride and so on are all indicators, all provide information on the (possible) fate of our goals [1] [57] [25] [28] [46]. As mentioned above, this is a particular type of information: immediate, global, not based on argument, far removed from analytical belief and from reasoning about the causes of success or failure. At the same time, however, it is precisely because of their immediacy and holistic nature that they constitute extremely effective signals and lead to equally immediate behavioural reactions (flight/approach).

iii. *The emotions activate goals.* The emotions not only monitor but also activate goals (the “reactions” referred to above). For example, in the case of envy (see 2.1), the “monitored” goal is “not to have less power” than the envied party, whereas the activated goal is to wish/do injury to the other party. In the case of guilt, the monitored goal is to respect shared rules or not to cause injury, while the activated goal is to make good the injury inflicted.

The activation of goals through emotions possesses very different characteristics from the activation of goals through beliefs. In cognitive activation [8], a goal *S1* cannot be generated by a belief. Instead, the belief *activates S1*, which was already represented in the mind, and can generate a sub-goal *S2* that serves the purpose of achieving *S1*. When I learn that the water will be cut off tomorrow, this activates the goal of “having water”, which interacts with the belief in question and activates the sub-goal of “laying in supplies”. Cognitive activation is closely related to the mechanisms of planning and of reasoning with respect to goals, means and conditions.

On the contrary, goals directly activated by emotions or “impulses” are “irrational” in that there does not exist a “logical” nexus of planning (i.e. in the mind of the agent, whereas its existence can be plausibly assumed outside the mind at the functional level) between activating conditions (including the monitored goal) and activated goal. To return to the example of envy, doing injury to the envied party is not a way (or rather is not represented as a way) of “not having less power” than him, or of gaining possession of a particular object of envy (motorcycle, promotion, holiday). In the “emotive” activation of goals, a belief (“Peter is better than me” or, to descend to the level of more elementary emotions like fear or disgust, “There’s a rat in the cellar”) directly “produces” the goal (“of wishing/making Peter suffer”, “of avoiding the rat”) without the mediation of evaluative knowledge and planning.

In the goals activated by emotion (which can range from specific goals to more general goals further removed from behavioural sequences), we basically resolve what is termed the *conative* component of the emotion, its tendency towards action.

Emotion and Motivation

In our view this is the main relationship between emotion and motivation: *emotion activates goals* (more or less important or urgent, and more or less close to an action sequence ready to be executed) and this can affect the commitment to previous intentions or stop a planned or executed behaviour. It is important to have clear that neither motivation can be reduced to emotions, nor vice versa. On the one side, emotions have several functions that are not depending on their motivational or conative aspect (for example affecting memory, or informing about beliefs and values); on the other side, there are motivational mechanisms that (at least in principle) are not necessarily based on emotions, like calculated preferences, planning, “cold” goals activation and pursuit, etc.

Cognitive Ingredients and Intension

Returning to the intensional aspect of emotion, we can now see its connection with the analysis of emotion. The *object* of an emotion appears to come from its activating beliefs, and especially from those to which the cause of the emotion is attributed. It comes from beliefs regarding the failure or attainment of a goal. Therefore it is also the “object” of a monitored goal. If I envy *y* for *i*, it means that I wanted *i*, that I do not think I (can) possess *i*, whereas *y* does, and that I feel ill-will for this reason. If I feel guilt about *i*, it means that *i* is an injury or injurious action that I regret having performed and that I relate my suffering to this responsibility.

The *target* of the emotion (in the case of social emotions) is connected with the goal activated. This is a social goal (to injure, to aid, to avoid) whose target is *y*. But this goal is connected with beliefs of which *y* is the object, which activate and justify the goal directed towards him: *y* is the person who I believe/imagine has more power than me (in the case of envy), or the person who I think I have injured (in the case of guilt), and so on.

3. Cognitive Ingredients: Envy, Shame, Trust

Let me give in this chapter a very brief and incomplete analysis of the most important cognitive components of envy and shame. I rely on the intuition of the reader since I cannot extensively justify this analysis here.

3.1 Necessary Beliefs and Goals for *Envy*

I consider envy in its strong and bad sense, as a serious social resentment and hostility (not simply as a desire).

(Envy x y p/i)

x envies y for p: having i

This necessarily presupposes that x

$$\frac{(\text{Bel } x (\text{Has } y \text{ i}))}{p}$$

- (Goal x (Has x i))

$$(\text{Bel } x (\text{NOT } (\text{Has } x i)))$$

However, not only *x* believes that he has not *i*, but he believes that he *cannot* have it (in time). This belief about lacking power is crucial for orienting the subject towards envy rather than towards emulation. Thus there is a power comparison where *x* is inferior

Power Comparison

$$(\text{Bel } x \ (\text{Can } y \ p))$$
$$(\text{Bel } x \ (\text{NOT} \ (\text{Can } x \ p)))$$
$$(\text{Bel } x (\text{More_Powerful } y \text{ } x \text{ as for } p)))$$

I assume that the envious person is not suffering only for not having i , but mainly for his inferiority: i.e. another more important (social) goal is frustrated. I assume that this is the real concern monitored by this emotion (not the specific lack of power).

Threatened Goal:

- (Goal x (NOT (More_Powerful y x ..)))
not having less power than the others

social hierarchy

Etc.

Suffering for both:

- (Goal x (Has x i)) x suffers for not having i
- (Goal x (NOT (More_Powerful y x ..))) x suffers for being inferior

This explains why the *goal activated* in/by envy is a social one and is a form of aggression, hostility towards the other, wishing to the other some misfortune, some loss of power.

You cannot really feel envy

- if you do not *believe* that y has i, and

The Meaning of Shame Expression

In our view [16] blushing has a specific meaning; it is a signal with a specific communicative function, although it bears a non-intentional, and even counter-voluntary, communication. Blush *communicates some crucial aspects of the mental state involved in shame*, namely the individual sensitiveness to others' judgement and, at the same time, his sincere sharing of their values and his suffering for being inadequate. The blusher is somehow "saying" he knows, cares and fear the others' evaluation, and he agrees about the standards and values; and he also communicates his sorrow over any possible faults or inadequacies of his, thus performing at the same time an acknowledgement, a confession and an apology aimed at inhibiting the others' moralistic aggression or avoiding social ostracism. The fact that blushing cannot be simulated guarantees the group about the sincerity of all this message and the internalisation of social values. Moreover, the meaning of blushing is perfectly coherent with a sort of non-verbal "discourse" whose components are the others typical shame signals and behaviours: the posture and movements of eyes and head. Those communicative (involuntary) acts converge towards a unique, final goal: inhibiting possible group aggression over an alleged violation of some value. Lowering the eyes communicates that the subject is giving up his power over the evaluator and accept some sort of subordination (and inferiority). The head lowering and the diminishing posture aims at the goal of looking smaller, and at giving up facing the others. The individual does not oppose to the group, accepts and cares about their evaluation: he argues for his belonging to the group, and at the same time he informs he is aware of his fault and has been already punished.

3.3 Explicit, Evaluation-Based Trust Vs Implicit Affective Trust

3.3.1 A Cognitive Anatomy of Trust

Let us briefly introduce our cognitive analysis of trust (for a more complete presentation see [12] [13] [14]).

In our model we specify which structure of beliefs and goals characterise x 's trust in another agent y .

Beliefs on Which Trust is Based

First, one trusts another only relatively to a goal, i.e. for something s/he wants to achieve, that s/he desires. If x does not have goals, she cannot really decide, nor care about something (welfare): she cannot subjectively "trust" somebody.

Second, at this level (see 2.3.2), trust itself *consists of* beliefs. Trust basically is a *mental state*, a complex mental *attitude* of an agent x towards another agent y about the behaviour/action a relevant for the result (goal) g .

- x is the *relying agent, who feels trust (trustor)*, it is a cognitive agent endowed with internal explicit goals and beliefs;
- y is the agent or entity which is trusted (*trustee*); y is not necessarily a cognitive agent (in this paper, however, we will consider only cognitive agents, i.e. social trust). So
- x trusts y "about" g/a (where g is a specific world state, and a is an action that produces that world state g) and "for" g/a ; x trusts also "that" g will be true.

Since y 's action is useful to x , and x is relying on it, this means that x is "delegating" some action/goal in her own plan to y . This is the strict relation between trust and reliance or delegation. *Trust is the mental counter-part of reliance.*

We summarize the main beliefs in our model:

1. "Competence" Belief: a *positive evaluation* of *y* is necessary, *x* should believe that *y* is useful for this goal of hers, that *y* can produce/provide the expected result, that *y* can play such a role in her plan/action, that *y* has some function.
2. "Disposition" Belief: Moreover, *x* should believe that *y* is not only able to perform that action/task, but *y* will actually do what *x* needs (predictability).
3. Fulfilment Belief: *x* believes that *g* will be achieved (thanks to *y* in this case). This is the "trust that" *g*.

When *y* is a cognitive agent these beliefs need to be articulated in and supported by more precise beliefs about *y*'s mind and/or personality: this makes *y*'s behaviour *predictable*. In particular, predictability or *y*'s disposition should be supported by:

4. Willingness Belief: *x* believes that *y* has decided and intends to do *a*. In fact for this kind of agent to do something, it must intend to do it. So trust requires modelling the mind of the other.
5. Persistence Belief: *x* should also believe that *y* is stable enough in his intentions, that *y* has no serious conflicts about *a* (otherwise *y* might change his mind), or that *y* is not unpredictable by character, etc.

While *y*'s competence might be complemented by:

6. Self-confidence Belief: *x* should also believe that *y* knows that *y* can do *a*. Thus *y* is self-confident. It is difficult to trust someone who does not trust himself!

We can say that trust is a set of mental attitudes characterizing the mind of a "delegating" agent, who prefers another agent doing the action; *y* is a cognitive agent, so *x* believes that *y* *intends to do* the action and *y* *will persist* in this.

Internal versus External Attribution of Trust

We should also distinguish between trust 'in' someone or something that has to act and produce a given performance thanks to its *internal* characteristics, and the global trust in the global event or process and its result which is also affected by external factors like opportunities and interferences.

Trust *in y* (for example, 'social trust' in strict sense) seems to consist in the two first prototypical beliefs/evaluations we identified as the basis for reliance: *ability/competence*, and *disposition*. Evaluation of *opportunities* is not really an evaluation about *y* (at most the belief about its ability to recognize, exploit and create opportunities is part of our trust 'in' *y*). We should also add an evaluation about the probability and consistence of obstacles, adversities, and interferences.

We will call this part of the global trust (the trust 'in' *y* relative to its internal powers - both motivational powers and competential powers) *internal trust*.

This distinction between internal versus external attribution is important for several reasons:

- To better capture the meaning of trust in several common sense and social science uses.
- To understand the precise role of that nucleus of trust that we could describe in terms of "unharmfulness", sense of safety, perception of goodwill.
- To better understand why trust cannot be simply reduced to and replaced by a probability or risk measure, like in the economic or game-theoretic approach to trust [14].

Trust can be said to consist of or rather to (either implicitly or explicitly) imply, the *subjective probability* of the successful performance of a given behaviour *a*, and it is on the basis of this subjective perception/evaluation of risk and opportunity that the agent decides to rely or not, to bet or not on *y*. However, the probability index is *based on, derived from those beliefs* and evaluations. In other terms, the global, final probability of the realisation of the goal *g*, i.e. of the successful performance

of a , should be decomposed into the probability of y performing the action well (that derives from the probability of willingness, persistence, engagement, competence: *internal attribution*) and the probability of having the appropriate conditions (opportunities and resources: *external attribution*) for the performance and for its success, and of not having interferences and adversities (*external attribution*).

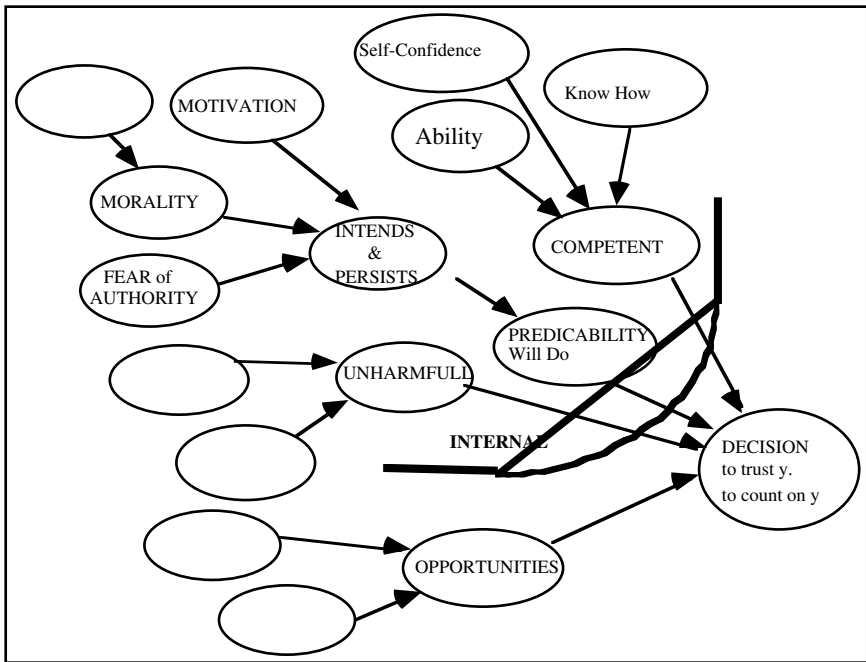
Degrees of Trust

The idea that trust is scalable is common (in common sense, in social sciences, in AI). However, since no real definition and cognitive characterisation of trust is given, the quantification of trust is quite *ad hoc* and arbitrary, and the introduction of this notion or predicate is semantically empty. On the contrary, we claim that there is a strong coherence between the cognitive definition of trust, its mental ingredients, and, on the one side, its value, on the other side, its social functions and its affective aspects. More precisely the latter are based on the former.

In our model we ground the degree of trust of x in y , in the cognitive components of x 's mental state of trust. More precisely, *the degree of trust is a function of the subjective certainty of the relevant beliefs*. We use the degree of trust to formalise a rational basis for the decision of relying and betting on y . Also we claim that the "quantitative" aspect of another basic ingredient is relevant: *the value or importance or utility of the goal g* . In sum,

- *the quantitative dimensions of trust are based on the quantitative dimensions of its cognitive constituents.*

For us trust is not an arbitrary index with an operational importance, without a real content, but it is based on the subjective certainty of the relevant beliefs.



3.3.2 Implicit and Affective Forms of Trust

Trust is also a "feeling", something that the agent "feels" towards another agent. It is *confidence* (similar to self-confidence) not a judgement. It can be not

argumentable and based on reasons or explicit experiences; it can be just "by default" or just "intuitive" and based on tacit knowledge and implicit learning.

At a primitive level (consider a baby) trust is something not express/ed/ible in words, not made of explicit beliefs about y's competence or reliability. It is a spontaneous, non reasonable or reasoned upon (non rational) reliance, and a feeling of confidence in a given environment or in a person. What is this kind or this facet of trust? [55]

Trust as a feeling is characterised by a sensation of "letting oneself go", of relaxing, a sort of confident surrender; there is an attenuation of the alerted and defensive attitude. Affective components of trust result in a felt freedom from anxiety and worry; x feels safe or even protected; there is no suspicion or hostility towards y which is appraised/felt as benevolent and reliable/able ("S/he will take care of..."). Towards a benevolent y we are benevolent, good-willing; towards a good/skilled y we are not aroused, alerted, cautious, wary (x could say: "I put myself into your hands"; while y will say "Let yourself go, do not resist, *trust* me"). Notice how these possible affective components of trust are coherent and compatible with our cognitive analysis of trust. However, they can also be independent of any judgement; they can be just the affective, dispositional consequence of an intuitive appraisal (cf. 4.2) and of learning. They can even be by default or just the result of lack of (bad) experiences, lack of negative evaluations.

Trust, Lack of Trust, and Mistrust

To understand part of this affective state it could be useful to distinguish between two different opposite of trust, two forms of "lack of trust": *distrust* and *mistrust*. There are two kinds of negative evaluation (cf. 4.1).

Evaluation of Inadequacy:

"x is not able to, is not good, apt, useful, adequate, ...for my goal";

(Bel x (Not(GOOD-FOR y p))) & (Goal x p) - (cf. 4.1)

Evaluation of harmfulness:

"x is good but for bad things (contrary my goal), it is noxious, dangerous"

(Bel x (GOOD-FOR y p)) & (Goal x (Not p))

Analogously there are *two forms of negative trust* or dis/mistrust, two opposites of trust; I propose to call them respectively "distrust" and "mistrust" (suspicion)³. At the explicit level, we have:

- **Negative Trust-1 or Distrust-** is when either x believes that x is not able, or that his/her behaviour is not predictable (s/he will not do a). x does not fear y; but it just *does not trust* y as for realising p or doing a.
- **Negative Trust-2 or Mistrust** is more than this. At the explicit level it is a paradoxical form of trust: x believes that y is able and willing to do something which is bad for x. x has negative expectations/evaluations about y's intentions and ability.

³ To be true, the natural language terms 'mistrust' or 'distrust' are ambiguous between the two forms. However, there is some connotation in one sense or in the other, which is the reason I adopted those terms as having a precise meaning.

Noticed that both imply **lack of trust**, i.e. the fact that x has not (enough) trust in y, as previously defined.

At the affective level this means either do not relax and quietly rely upon, or even it means suspicion, apprehension: x worries about y's action. S/he does not trust y in the sense that s/he is cautious and suspicious.⁴

Implicit Trust

Given these two different opposites of trust, one can identify an important implicit form of trust where x has neither specific and explicit positive evaluations (cf. 4.1) about y, nor has suspects and worries. S/he is just *without doubts, suspect and worries*, s/he naively relies upon y: not because of an explicit and reason based evaluation of y. S/he trusts by default and because s/he has no reasons to be cautious and to suspect, and then without any explicitly exam of whether y is able, willing, or dangerous. One could represent this attitude as the absence of Distrust and of Mistrust, of suspicion, and also of explicit positive evaluations and trust.

This implicit, passive, and spontaneous or naive form of trust consists of not having the typical trust beliefs, but also in *not having* negative ones, negative expectations: to be without any suspect and suspicion. Consider in fact that not having a given belief (I do not believe that it is raining) is a very different cognitive state that having the opposite belief (I believe that it is not raining).

As not trusting y for p is not the same of expecting harm from y, analogously not dis/mistrusting, do not worrying is not the same of positively believing that y is capable and good-willing. However, this lack of dis/mistrust can be sufficient for relying upon y.

The behaviour of trust consists in this weak form in the absence of cautions, of controls, of any search for evidence for evaluation, and in the absence of a true "decision" about trusting or not y.

Only after some negative unexpected experience, this kind of trust is damaged. Who uses explicit, evaluation-based trust, based on evidence, is no longer naive: s/he has already considered the situation as problematic, S/he has some doubt. There is on the contrary a form of trust without and before any question like: "Can/should I trust y?".⁵

⁴ One might (by paradoxically using a predicate *Trust* without the goal in its definition) represent Positive trust, Lack of trust, and Negative trust as follows:

(Trust y x p) & (Goal y p) Positive Trust

(Not (Trust y x p)) & (Goal y p) Lack of Trust

(Trust y x p) & (Goal y (Not p)) Mistrust

⁵ It is important also to distinguish between *uncertainty* (the fact that we do not have complete evidence of our positive evaluation of -trust in- y, we are not hundred % sure of our beliefs), that make y's behaviour (and results) not completely subjectively predictable; from the actual presence of contrasting, negative evaluations and expectations. The absence of a belief is a mental state significantly different from the presence of the negative belief, with completely different consequences at the reasoning and at the pragmatic level. When x has positive evaluations of y, and does not have any negative (pertinent) evaluation, although this positive evaluation leaves some room to ignorance and uncertainty, this is very different from a situation where x has negative beliefs about y which make y 'ambivalent' (attractive and repulsive, positive and negative, at the same time) and destroy x's 'trust in' y, his trustworthiness. Non-ambivalent although uncertain

Let me conclude this very preliminary exploration of such a hard topic as follow: the lack of explicit trust seems to cover three different mental states:

- insufficient trust (x does not estimate enough y to count on him, she has some negative evaluation on y) or distrust
- mistrust (x worries about y)
- implicit trust, be it either spontaneous, naive and by default (lack of suspect), or be it automatic and procedural, just based on previous positive experience and learning.

4. Emotions without “Feeling”?

It should be time now for cognitive models to go beyond the functional approach to emotions. The title of the well-known paper by Frijda and Swagerman [27] was “Can computer feel?”, but immediately the authors clarify that:

“The question is ambiguous. The word “feeling” refers to a kind of conscious experience and to a process of affective appraisal and response... The paper discusses the second meaning of “feeling”. Its main theme can be rephrased as: Can computers *have* emotions?”.

Compared with the title of the paper this is a bit disappointing. Can cognitive science provide any theory of the most relevant sense of “feeling”, the subjective, phenomenic experience? Or is this beyond the capability of the cognitive science paradigm? In my view, Frijda’s claim that “what is interesting in the phenomena that make one use concepts like ‘emotions’ and ‘emotional’ is not primarily subjective experience” (p.236), should be considered now a bit obsolete. One should not abandon either the functionalist approach to emotions or the cognitive analysis of emotions, but one should try to extend the model to cover or to integrate the other very relevant aspects. So the question “Can computer feel?” should be taken seriously. I claim that computers can feel if they have a body (and a brain) not simply a hardware: a real body including self-perception and internal reactions (either peripheral or central); and these somatic reactions should be related -by attributional representations or by association- to mental representations (beliefs, goals, etc.).

It is important to understand that the problem is not only to go beyond a functionalist analysis of emotions to integrate other aspects, but the problem is that

evaluation is very different from ambivalent evaluation. Thus, we have to distinguish between two types of ‘unharmfulness’ ‘safety’ ‘there is nothing to worry’ etc.: the implicit and the explicit one.

Implicit unharmfulness simply consists in the absence of suspects, doubts, reasons to worry, diffidence, no perceived threats; some sort of ‘by default’ naive and non-argumentable confidence. I do not have reasons to doubt of y’s pro-attitude (active or passive adoption), I do not have negative beliefs about this.

Explicit unharmfulness consists of explicit beliefs about the fact that ‘I have nothing to worry from y’.

Both the implicit or explicit unharmfulness can be based on other beliefs about y, like “He is a friend of mine”, “I’m likeable”, “I feel his positive emotional disposition” (empathy), “he is honest and respectful of norms and promises”, “he fears me enough”, ... and also “he trusts me and relies on me”. This unharmfulness perception and then trust in y based on y’s trust in x, is important for the circular dynamics of trust and to explain how trust can create trust.

- *any functional explanation is incomplete if ignores the subjective facet of emotions.*

The real problem is precisely the function of the internal perception, of the “feeling” of the bodily peripheral reactions and of the central response. *Since a reactive system can do the job of an emotional system, why do we need emotions?* why do we need a system which perceives its own reactions? what is the role of this self-perception in the adaptive process?

Let me clarify a bit this point. With a few exceptions (for ex. [43]) the dominant AI position about emotions remains that enounced by Simon [48] who explains their function in terms of operating system interrupts that prompt one processing activity to be replaced by another of higher priority, i.e. in terms of a reactive goal-directed system in an unpredictable environment. As Sloman and Croucher observe, the need to cope with a changing and partly unpredictable world makes it very likely that any intelligent system with multiple motives and limited powers will have emotions [50]. I believe that this view is basically correct, rather topical (ex. [42]) but seriously incomplete. This function is necessary to explain emotions but is not sufficient; and precisely AI and ALife show this. In fact, to deal with this kind of functionality a good reactive multi-task system able to focus attention or memory and to activate or inhibit goals and actions would be enough. Current models of affective computing simply model the emotional behaviour/expression and the cognitive-reactivity function. Consider for ex. Picard's nice description of fear in a robot:

"In its usual, nonemotional state, the robot peruses the planet, gathering data, analyzing it, and communicating its results back to earth. At one point, however, the robot senses that it has been physically damaged and changes to a new internal state, perhaps named 'fear'. In this new state it behaves differently, quickly reallocating its resources to drive its perceptual sensors and provide extra power to its motor system to let it move rapidly away from the source of danger. However, as long as the robot remains in a state of fear, it has insufficient resources to perform its data analysis (like human beings who can't concentrate on a task when they are in danger). The robot communication priorities, ceasing to be scientific, put out a call for help." [44]⁶

What is lacking in this characterisation of fear? just the most typical "emotional" aspect: feeling.

Feeling is a broader notion: we can feel a lot of things that are not emotions (needs, for ex.). However, feeling is a kernel component of emotion: if we cannot feel x, we should/could doubt that x is an emotion [39]. As I said, this puts out a serious question: since we can account for emotional functioning without modelling feeling, since a reactive change of the internal state, cognitive processing, and behaviour is enough, *why is feeling such a crucial component of human (and animal) emotions?* Is it a mere epiphenomenon lacking any causal function in the process? Or which is its function and its reason?

Emotions, it is true, provide reactivity to the agent, and enter the control and the commitment of the system. More precisely, emotions are -as we saw- one of the sources of our goals, they are one of the mechanisms of what I call "cognitive reactivity" [8]:

a) the reactive activation not of a reflex or an action but of a goal (to be planned, in case), and

b) an activation based not on a stimulus/perception but on a belief (even internally generated by inference).

The emotionally activated goals can also bypass decision-making (or have high priority), and interrupt planning or execution. All this is rather obvious.

⁶ A similar view in (Webster 1997) about depression in robots.

However, we already have this kind-level of reactivity for example in BDI agents, or in other architecture more biologically inspired. Thus, we don't really need emotions for doing this, nor there is any reasons to call "emotions" such a reactive mechanisms, since they do not capture what is special in emotions. Is there something more than reactivity in emotions? something we don't have yet in our cognitive-reactive agents, and that could be important for their efficiency/survival? Consider a BDI architecture (Georgeff; Lansky; Pollack; Rao; see [29]): the agent is *situated* in an evolving environment and is able to modify its behaviour -opportunistically and reactively- on the basis of the information arriving from the environment relative to success or failure of actions, to changes in the world, erroneous predictions, etc. This is rather "adaptive".

But living organisms have also an "internal environment" they receive information from and react to. Let call "internal states" the states of this internal environment (the body and its self-perception) which are different from mental/cognitive states (I mean the state of the agent's beliefs, goals, intentions, etc..).

These internal states can evolve in time following rules that are independent of the evolution of mental states. Consider for example the level of energy (battery) at disposal for actions: this level will decrease with the activity of the agent and/or with time. The system perceive this (has an information about this), and its preferences can change on such a basis.

These internal states evolve in time following rules that are also independent of the evolution of the external environment. However, these internal states also evolve on the basis of external stimuli: reactively (this is specially true for emotions), and we perceive such "alteration" of our internal state, we *perceive* our bodily "response" to the environment. For example the arousal reaction.

Which is the role of this "internal environment" and of the information about it? Let me say that this role is "dispositional": the status of the internal environment change the "disposition" of the agent toward the external environment. "Disposition" of course is an awful psychological, very vague and ambiguous term. I do not mean here goal activation, the orientation toward a given behaviour. I prefer to call this latter aspect of emotion: "conative" or "motivational". Thus, let me try to characterize the "disposition" saying that the internal state changes and "tunes" all the information processing, and the cognitive and motivational processing of the system.

For example, the thresholds of our perceptual apparatus or of our reactions are modified; the rapidity of certain kind of processing (or of all the processes) is improved; resources of processing are allocated on certain levels or tasks; attention is oriented to certain area or stimuli; some perceptual or semantic aspects are activated; some goals are more ready to be activated than other; they will be activated with a different urgency or force; then also preferences will be different; the same for possible alternative behaviour for the same purpose; the credibility of certain sources or beliefs will change; memory is affected too; etc.

In other words: all the processing of perception, beliefs, and goals is affected by a given internal state and its impact on the BDI architecture.

In other words: the response of the agent to the same external event or stimuli -being equal its knowledge, its ability, its potential goals (desires/motives)- will be different depending on its internal state (and *the information it has about*).

This is true not only for emotions but even for simpler states like hunger or the sexual drive. It is well known that animals that react to a given sexual stimulus with a fixed pattern of behaviour (reflex) do not react always and always with the same probability, easiness, and intensity. It depends on the internal level of the sexual drive. It is well known that we are more leaning to see (perceive or misperceive) food or things associated with food (ex. food related words) when we

are hungry, and that all our preferences, priorities, attention and memory processes, etc. change. This is even more true for emotions in the strict sense.

Emotions also show some differences (if compared with hunger, etc.), of course: the internal emotional status can be created just by external stimuli or *by beliefs*: normally they are not internally/physiologically generated. Thus, in some sense we react (at a cognitive level: the dispositional impact of the emotion) to a sub-cognitive reaction: *we react to our reactions*.

Thus, again: to stress the importance of emotion for autonomy, reactivity and adaptation in a dynamic environment, or for dealing with multiple goals [42] is not enough. What is specific and special in emotions as reactive control systems? why shouldn't other multi-task, reactive and situated mechanisms be enough? What is the role of feeling, of "sensing" the internal environment and the automatic bodily reactions in this reactive control and adjustment to the stressing external environment?

I believe that computational models of emotions should answer precisely this question, which they unavoidably elicit.

As I already said elsewhere [9] I believe that the main functions of the *feeling component* in emotion, i.e. of the fact that the robot should sense those changes of its internal state and of its behaviour (energy allocation, disturbance and distraction from the task, fast movement of avoidance, etc.) are the following ones:

- a) felt emotional internal states work as *drives* [41] [6] [5] to be satisfied, i.e. to be brought back to the equilibrium (homeostasis) through action; Mower [38] postulates that in learning the animal learns precisely what behavior serves to alleviate the emotion associated to a given stimulus;
- b) felt emotional internal states work as positive or negative *internal reinforcements for learning* (they will be associated to the episode and change the probability of the reproduction of the same behaviour);⁷
- c) last but not least, as we will see felt emotional internal states associated to and aroused by a given scenario constitute its immediate, unreasoned, non-declarative *appraisal* (to be distinguished from a cognitive evaluation - cf. 4.1).

The cognitive dominant paradigm cannot neglect any longer the necessity for modelling subjective experiences and feelings. The relation with a body seems to be crucial: beliefs, goals, and other mental (declarative) ingredients do not seem to be enough. Nevertheless, *beliefs, goals, expectations are necessary also for modelling emotions*, at least complex social emotions. Also a better and convincing functionalist analysis of emotions (see also [19]) requires precisely to explain the functional role of "feeling": cognitive appraisal, modification of attention and of cognitive processes, reactive changes of goal priorities, are not sufficient.

Let's now examine a bit more carefully the two natural systems of valence and how they are related to emotions and interaction.

5. The Dual Nature of Valence: Cognitive Evaluations *versus* Intuitive Appraisal

*Heart has its own reasons
which are unknown to Reason*
Pascal

⁷ I assume, following a long tradition on emotional learning, that in general positive and negative emotions are reinforcers; but notice that this does neither imply that we act in order to feel the emotion, which is not necessarily motivating us (it can be expected without being intended); nor that only pleasure and pain, or emotions, are rewarding (Staats, 1990). See also Bozinovski, S. and Bozinovska, L. (1998) .

There are at least two kinds of appreciation of valence of events, situations and entities; two kinds of "evaluation" in a broad sense.

- a)** A declarative or explicit form of evaluation, that contains a judgement of a means-end link, frequently supported by some reason for this judgement, relative to some "quality" or standard satisfaction.

This is a reason-based evaluation that can be discussed, explained, argued upon. Also the goal of having/using the well-evaluated entity (which is the declarative equivalent of "attraction") can be "justified". This is the classical approach to values (moral or ethical) that is synthesised by the "motto" (of Aristotelian spirit):

"it is pleasant/ we like it, because it is good/beautiful"

- b)** A non-rational (but adaptive) evaluation, not based on justifiable reasons; a mere "appraisal", which is just based on associative learning and memory.

In my view, in the psychological literature on emotions, in particular in the very important and rich literature on emotions as based on a cognitive appraisal of the situation [25] [26] [1] [47], there is a systematic and dangerous confusion between these two kinds of "evaluation" (also in Damasio). Incoherent terms and properties are attributed indifferently to the term "appraisal" or "evaluation". This fundamental forerunner and component of the emotion is characterised -at the same time- as "cognitive", "intuitive", "immediate", "unconscious", implying also inferences and predictions, etc. I propose [37] to distinguish between "appraisal" - that should be the unconscious or automatic, implicit, an intuitive orientation towards what is good and what is bad for the organism- and "evaluation". I reserve this term for the cognitive judgements relative to what is good or bad for p (and why).

5.1 Evaluations

Let me -very shortly - characterise what kind of mental object is an evaluation, how it is reducible to its elementary components (beliefs and goals), how it works (for a well argued and extended analysis of this see [37].

Our claims are that:

- an evaluation is a kind of *belief* concerning the power (hence the usefulness) a given entity (be it an object, organism, agent, institution, etc.) or state of the world is endowed with in relation to a certain goal; evaluations are closely linked to goals: they are beliefs arising from goals, and that give rise to goals; evaluations play a crucial role both in cognition (problem solving) and in social interaction;
- a value is a special kind of evaluation; while an evaluation implies a means-end relationship (an entity or state of the world is assumed to be "good" or "bad" for a goal p) a value is a "means" for an unspecified goal, or class of goals, and turns into something "good" in itself; values show an even closer relationship with goals, and in particular norms, by virtue of the absolute character of this special kind of evaluations; values serve important functions, both cognitive and social, which result from their being a borderline mental object, between evaluations and goals: values' kinship with absolute imperatives in fact favours the social function, while their cognitive functions are made possible by their evaluative features.

Let's put aside values and see a bit more precisely cognitive evaluations.

We define an evaluation of an entity x as a *belief of an evaluating agent e about x 's usefulness with regard to a goal p* . If for instance x is a pair of scissors and e

believes -- from direct experience, inference, or someone else's communication -- that it is good for cutting a piece of cloth, in so doing e is evaluating the scissors with regard to that goal. We might represent an evaluation (more precisely a 'positive' evaluation) as follows:

(BEL e (GOOD-FOR x p))

where x denotes an entity variable (i.e., an object of any kind: physical object, organism, agent, etc.), e is an agent variable and p is a well-formed formula representing a state of the world; the predicate (GOOD-FOR x p) means that x is a means for p , which is what e BELIEVE. GOOD-FOR has a very broad semantics: it merely expresses that x (or q) is useful for making p true; x may either directly realize p , or cause p to be realized, or favour that p be realized. As already observed, GOOD-FOR expresses a very broad means-end relationship. Evaluations are a special kind of beliefs, characterized by a strict relationship with action, by virtue of their link with goals.

Evaluations imply goals by definition, in that the latter are a necessary component of evaluations, namely, the second argument of the GOOD-FOR predicate. From a more "substantialist" perspective, evaluations imply goals in the sense that they *originate* from them: it is the existence of some goal p (either e 's or someone else's) that make the word good or bad, justifies and motivates both the search for a means x to achieve it, and the belief that x is (not) GOOD-FOR p . Goals and evaluations endow objects and people with 'qualities' and 'faults'.

The relationship between evaluations and goals is even closer, because evaluations not only imply goals, but also can *generate* them. In fact, if e believes x is good for some goal, and e has that goal, e is also likely to want (possess, use) x . So there is a rule of "goal generation" which might be expressed as follows: if e believes something x to be a means for e 's goal p , e comes to have the goal (USE e x) of exploiting the means x .

Evaluations, that is knowledge about "what is good for what", and "why", play a crucial role in all the cognitive activities which are based upon symbolic and explicit representations, reasoning and deliberation. For example in problem solving and decision making the particular advantage offered by evaluative knowledge is precisely a preliminary relationship established between descriptive knowledge and goals, in terms of beliefs about "what is good for what", derived from either one's experience about problems solved in the past, or one's reasoning and inferences (think for instance of evaluation by standard), or others' communication. Evaluations make such a relationship explicit; they fill the gap between knowledge and goals, by "reinterpreting" the properties, qualities, and characteristics of objects and situations in terms of *means* for the system's (potential or actual) goals. The cognitive network ceases to be neutral and becomes "polarized" toward goals, that is ready for problem solving and decision making.

In a cognitive agent preferences can be internally represented both at the procedural and at the declarative (propositional) level.

- Having a "procedural preference" means that, at a given level of their processing, a system's goals present different degrees or indexes of activation, priority, weight, value, importance (or whatever), that in fact create some rank order among them, which will be followed by some choice/selection procedure.
- Having a "declarative preference" means that the system is endowed with an explicit belief like: " x is better than y (for goal p)". In particular, three types of beliefs are relevant for preferences: (a) simple evaluations, that is beliefs about how good/useful/apt/powerful are certain entities relative to a given goal (" x is very useful for p "; " y is quite insufficient for p "); (b) comparative evaluations like " x is better than y for p "; (c) reflective preference statements, of the kind "I prefer x to y (for p)". Generally, (b) are based on (a), and (c) are based on (b).

Both procedural and declarative preferences can coexist in a human mind (and would be of some use in artificial minds too), and each level of preference representation, though having its own mechanisms of reasoning, is translatable into the other. One can derive a "weight" from the declarative evaluations and their arguments, and vice versa, one can explicitly express (as beliefs) some priority of attractiveness, urgency, activation, or whatever.

However, being able to *deliberate*, that is, to choose an alternative on the grounds of explicit evaluations concerning the "goodness" of the various options, and being capable of reasoning aiming at supporting such judgements add further advantages to the mere fact of making choices. In these cases, in fact, the system can *justify* its choices, as well as *modify* the "values" at stake through reasoning. Moreover, it is liable to persuasion, that is, it can modify its preferences on the grounds of the evaluations conveyed by others.

We interact with people on the basis of the image and trust we have of them, i.e. on the basis of our evaluations of them: this define their 'value' and reputation. And also social hierarchies are just the resultant of the evaluations that the individuals and the groups receive the one from the others.

Evaluations and Emotions

Given this "cold" view of evaluation ("cold" if compared it with others' - e.g., [34], what is the relationship between evaluation and emotion? As we claim in [37]:

- ***Evaluations do not necessarily imply emotions***

No doubt many evaluations show some emotional feature. For instance, if I believe a certain food, book, person, and so on, to be "good", I will be likely to feel attracted to it (or him or her). But evaluations and emotions are not necessarily associated with each other, because not any belief about the goodness or badness of something necessarily implies or induces an emotion or an attraction/rejection with regard to that "something". There also exist "cold" evaluations: if, for instance, I believe that John is a good typist, I will not necessarily feel attracted to him. Evaluations are luckily to have emotional consequences if they:

- i) are about our own goals (e, the evaluator, is = to x, the goal owner);
- ii) these goals are currently active;
- iii) they are important goals.

- ***Emotions do not necessarily imply evaluations***

One may view attraction or rejection for some x as a (possible) consequence of an evaluation; so, in this case the emotion "implies" an evaluation in the sense we have just considered. On the other hand, however, one may view attraction or rejection *per se* as forms of evaluation of the "attractive" or "repulsive" object. In the latter case, we are dealing with a supposed identification: to say that an emotion implies an evaluation means to claim that the two actually coincide, which is still to be proved.

In fact, we view attraction and rejection as *pre-cognitive implicit evaluation that we call "appraisal"*.

5.2 Appraisal

We assume that a positive or negative emotional response can be associated with some stimulus. The automatic activation of this associated internal response (in Damasio's terms, a "somatic marker"; [20]) is the "appraisal" of the stimulus postulated by several theories of emotions [1] [25] [32] [51]. The associated negative or positive emotion makes the situation bad or good, unpleasant or pleasant, and we dislike or we like it.

- “Appraisal “consists of an automatic association (conscious or unconscious) of an internal affective response/state either pleasant or unpleasant, either attractive or repulsive, etc., to the appraised stimulus or representation.

It does not consist in a judgement of appropriateness or capability - possibly supported by additional justifications- ; on the contrary, it just consists in a subjective positive or negative experience/feeling associated to the stimulus or to the mental representation, usually previously conditioned to it in similar circumstances, and now retrieved.

This gives us a completely different “philosophy” of valence and value: now the “motto” is the other way around -in Spinoza’s spirit-

“It is good/beautiful what we like/what is pleasant”

As a cognitive evaluation of x is likely to give rise to some goal (if the evaluator e believes something x to be a means for e ’s goal p , e comes to have the goal q , of acquiring and using the means x - see [37]), also the emotional appraisal of x gives rise to a goal: it activates a very general goal linked to the emotional reaction. This is the *conative* aspect of emotional appraisal. Positive appraisal activates an “approach goal” (“to be close to x ; to have x ”), while negative appraisal activates a generic avoidance goal (“not to be close to x ; to avoid x ”).

We consider these sub-symbolic, implicit forms of “evaluation” as evolutionary forerunners of cognitive evaluations. Thus we believe the answer to the question “do emotions imply evaluations” depends on the level of analysis addressed.

It seems to me that there are several partially independent scientific traditions that can now contribute to give us an operational and clear model of what appraisal is and how account for its primary, intuitive, reactive, automatic, etc. character and its link with ‘feeling’. We do not claim that such a model already exists, but simply that it can be guessed, because those approaches provides converging evidences and converging models about mechanisms strongly candidates to implement the psychological notion of appraisal as distinguished both

- from generic primary stages of information processing of the stimulus ⁸, and
- from high level declarative cognitive evaluations.

I refer to studies from some pavlovian and behaviourist tradition (ex. [52]), studies about the affective primary automatic response (ex. [3]), and Damasio’s theory of “somatic markers”. I cannot discuss this literature here.

Elements for a Model of Appraisal

It seems to me that those approaches characterise process that partially overlap and partially can coexist, and that they can provide a quite clear idea of how to define Appraisal as clearly distinct from cognitive evaluations.

Main feature of such a model should be the following ones.

1. Appraisal is an *associated, conditioned somatic response* which has a central component and involves pleasure/displeasure, attraction/repulsion. Where

⁸ Consider for example the pattern matching process and the recognition of a key-stimulus or releaser in animal. For example the recognition of the right configuration of spots on seagull beak elicits the pecking behaviour in the offspring. We do not consider this information processing as such as an ‘appraisal’ or an ‘evaluation’ of the stimulus. Per se it does not imply any ‘valence’ any effective reaction or any judgement of goodness. The stimulus is not positive or negative, till we consider only its recognition as a pattern matching, and its unconditioned behavioural response. If we suppose that it elicits also some ‘emotional’ or ‘pleasure’ or ‘attractive’ response (which does not simply coincide with the behaviour) associated to it, in this case this is the appraisal response, and there is an appraisal of the stimulus, which gives it a subjective ‘valence’.

attraction/repulsion is not a motor behavior but just the preliminary, central and preparatory part of a motor response. And pleasure/displeasure is simply the activation of neural centres.

2. This associated response can be *merely central*, because also the somatic-emotional component can be reduce to its central trace (Damasio's *somatic markers*) and because emotions have a central response component which is fundamental [35] [52]. But of course this response can be also more complete involving overt motor or mussel responses or somatic emotional reactions.

3. This associate response is *automatic*, and *frequently unconscious*.

- Appraisal is a way of '*feeling*' *something*, thanks to its somatic (although central) nature.
- Appraisal *gives* '*valence*' to the stimulus because makes it attractive or repulsive, good or bad, pleasant or disagreeable.
- Appraisal has '*intensionality*' i.e. the association/activation makes what we feel "about" the stimulus, makes it nice or bad, fearful or attractive. It gives the stimulus that character that Wertheimer called 'physiognomic'. (How this happens, how the associated responses is 'ascribed to', 'attributed to', and 'characterises and colours' the stimulus; how it does not remain concurrent but dissociated is not clear -at least to me- and probably just the effect of a neural mechanism).

4. When it is a response just to the stimulus it is *very fast*, *primary*. It anticipates high level processing of the stimulus (like meaning retrieval) and even its recognition (it can be subliminal). In this sense the old Zajonc slogan 'preferences need no inferences' prove to be right (although not exclusive: there are preferences which are based on reasoning and inferences; and also emotions based on this).

5. There can be an analogous associative, conditioned, automatic *response to high level representations*: to beliefs, to hypothetical scenarios and decisions (like in Damasio⁹), to mental images, to goals, etc.

We have to change our usual view of cognitive 'layers', where association and conditioning are only relative to stimuli and behaviours, not to cognitive mental representations.

- **Any emotion as a response *implies an appraisal*** in the abovementioned sense.

It implies the elicitation of a central affective response involving pleasure/displeasure, attraction/repulsion, and central somatic markers if not peripheral reactions and sensations. This is what gives emotions their 'felt' character. (While, not all emotions presuppose or imply a cognitive evaluation of the circumstances).

5.3 Relationships between Appraisal and Evaluation

Evaluation and emotional appraisal have much in common: their function (Miceli and Castelfranchi, in press). Evaluations favor the acquisition of adequate means for one's goals, and the avoidance of useless or dangerous means, and precisely the same function can be attributed to emotions.

⁹ Although he seems not completely aware of this important feature of his theory, which makes it different from traditional conditioning theories.

More than that: emotions -- though they have traditionally been attributed the negative role of clouding and altering rational thought -- seem to help at least some kind of reasoning. In fact, they provide "nonconscious biases" that support processes of cognitive evaluation and reasoning [2], enabling for instance to choose an advantageous alternative before being able to explicitly evaluate it as advantageous ¹⁰.

However, all this should not prevent one from acknowledging the differences between emotional appraisal and cognitive evaluation, addressing the latter in their own right, and trying to establish their specific functions (Miceli and Castelfranchi, in preparation). For instance, in some context emotional appraisal by itself might prove insufficient for assuring adaptive responses, in that, the more changeable and complex the world becomes (because of the increasing number of goals and situations to deal with, and the complex relations among such goals and contexts), the more one is in need of analytical and flexible judgements about objects and events, rather than (or in addition to) more global and automatic reactions. In fact, evaluations allow to make subtle distinctions between similar (but not identical) goals and means, and to find out the right means for some new goal, never pursued in the past.

Moreover, evaluations allow to reason about means and goals, and to construct and transmit theories for explaining or predicting the outcome of behaviour.

Therefore, though *emotional appraisal can be conceived of as an evolutionary forerunner of cognitive evaluation* (as well as a valuable "support" for it), being an evolutionary "heir" does not imply maintaining the same nature as the forerunner; on the contrary, one might suppose that the same function has favored the development of different means, at different levels of complexity.

It is also important to consider that evaluation and appraisal about the same entity/event can co-occur, and give rise to *convergence* and enhancement of the valence, or to *conflicts*; in fact, either

- the means that we are rationally considering for our ends are associated to previous or imagined positive experiences); or
- what I believe to be the right thing to do frightens me; what I believe to be wrong to do attracts me.

(On this Damasio's model of the role of the somatic markers in decision making seems rather simplistic).

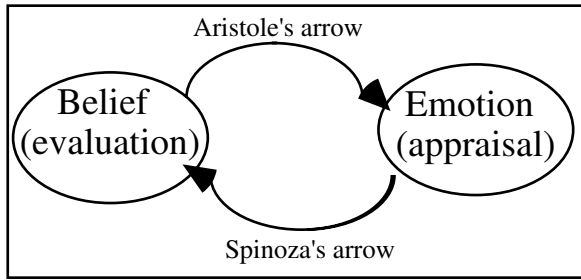
Evaluation and appraisal can also derive one from the other.

It is possible to verbalise, to translate a merely affective reaction towards B into a declarative appreciation. This is for example what happens to the subjects in the experiment by Bargh [3]. They do not realise that their evaluation is just a post hoc rationalisation of some arbitrary association they are not aware of.

¹⁰ A number of studies conducted by Damasio and his collaborators (e.g. Bechara, Damasio, Damasio, and Anderson 1994) have pointed to the crucial role of emotion in cognitive evaluation and decision making. Their patients with lesions of the ventromedial prefrontal cortex show emotional blunting as well as difficulties in making decisions, especially in real-life contexts. If compared with normal subjects, they do not show stress reactions (as measured, for instance, by skin conductance response) when trying to make choices in uncertain and risky contexts (e.g. a gambling task). The interesting fact is that such emotional reactions, displayed by the normal subjects especially before making a wrong choice (i.e. a kind of choice previously associated with some punishment), help them to avoid it, and to opt for a less risky alternative. Such a choice is made before reasoning over the pertinent beliefs, including cognitive evaluations about the game, its options, and the possible strategies of decision making.

Also the opposite path - from a cold evaluation to a hot appraisal - is possible; especially -as we saw- for personal, active, important goals, and in particular for felt kinds of goals like needs, desires, etc. [9].

This possible translation from one form to the other is very important also because it helps to explain a very well known vicious and irrational circle of our emotional life [24]. I mean the fact that we feel our emotional activation, what we feel towards B, as a possible evidence, *confirmation* of our beliefs that gave rise to that emotion itself. So for example we start with a belief that B can be dangerous, we predict possible harm, on such a basis we feel some fear, and then this fear (as an appraisal of B) 'feedbacks' on the beliefs and increases their certainty, i.e. confirms them; something like: "Since I'm afraid, I was right! it is dangerous" (which is not such a rational evidence; it is a case of self-fulfilling prophecy and of "motivated reasoning" [31]):



6. What Aspects of Emotion Are Pertinent in Interaction? Cognitive Ingredients? Feeling? Evaluation? Appraisal?

The need of giving more weight to emotional and personality aspects, in agents' behaviour modeling, is emerging as a new and promising research objective [30] [40]. A 'believable' behaviour would be shown by an agent which avoids answering to questions to which it is not able or willing to answer, is able to react to unexpected questions or attitudes of other agents, and assigns weights to its goals and beliefs according also to its emotional status [45]. Emotional and personality aspects of the mental state play a particularly relevant role in conflicts among agents: 'impatient, aggressive, polite, persistent, rude....' people show radically different ways of managing conflicts with others [11]. In general, emotion will play an important role in HC, in H-Agent and probably also in Agent-Agent interaction [40]. However, I will discuss on this topic only:

- the role of cognitive emotional stuff (mainly: evaluations);
- the role of the feeling component;
- the role of affective/appraisal reactions in interaction, compared with the role of true cognitive evaluations.

6.1 Are the Cognitive Aspects of Emotion Useful in Interaction?

Is there any advantage or necessity of **cognitively complex emotions** in HCI? I believe that in several cases a cognitive content of emotions will be useful and even necessary.

Let's consider for example the *recognition* of the other's emotions.

Emotional Communication: The "Meaning" of Faces

There is an *implicit and reactive* "meaning" or better functional social effect of the expression of emotions; but there is also some "*symbolic*" meaning, something that the emotion is communicating to us and that we explicitly know and understand.

When A recognises B's emotion this is in fact in order to have an "appropriate" reaction to /interaction with B, either immediately (during the emotional episode) or in the future: A learns about B through her/its emotional expression. But what is an "appropriate" reaction? To what it is "appropriate"? Or: what does A learn about B?

When A recognises an emotion he is not just recognising its bodily or behavioural signs. The recognition of an emotion is somewhat similar to the "recognition" of an intention or a plan from an observed action: it should precisely entail *the attribution to B of an internal state of mind* or subjective experience.

What A learns about B is mainly about her mind (her goals, concerns, beliefs, feeling and personality); and it is to B's mental content that A's reaction should be "appropriate".

If A recognises that B is ashamed and blushing (in a given situation) but he/it does not *understand* "why", i.e. what B is worrying about:

- whose negative judgement B is caring;
- what she/it considers a poor performance or a defect;
- why she/it believes that the other noticed or discovered this;
- what norm, value or expectations she/it believes to have violated; and so on;

A is *not really able to efficaciously interact* with B on that emotional episode. He/it cannot for example say appropriately:

"there is nothing wrong in doing/being.....", or "nobody will know about"; "why do you worry so much about y's opinion..."; "You should not be ashamed of ...; there is nothing bad in this; you should be proud of..."

A cannot take the appropriate decision for the future, for instance, avoiding to expose B to some situation or information that could be embarrassing for her.

This applies both to the recognition of the user's emotions, and to the recognition of the machine's (or believable agent's) emotion. In case of a machine's expression of emotion it will not be able to appropriately interact about this, because it will not be able for example to explain what and why it is blushing; what it believes (for example that some other agent saw it doing something, or that somebody can infer/find out that something is bad, and that it is responsible for that).

The same applies in emotion production. If the machine/agent must simply produce funny faces and expressions, a non-cognitive emotion is enough; but if it must be able to interact about this, in dialogue for example, there will be the above problem.

- *For an advanced emotional interaction the artificial entities should have an analytical/complex representation of the presupposed mental state (beliefs, goals).*

It is important to stress that not necessarily they must *have* this mental content to show the emotion. Knowledge about and the capability to "reason" about this mental stuff seem enough: like an alien observing and studying people on the earth and simulating emotions (and discussing about them).

However, if A has to simulate having a given goal, given beliefs, a given value or norm, etc. in order to efficaciously simulating of having a certain emotion, perhaps

it would be simpler to really build some sort of BDI mind in order to express emotions.¹¹

6.2 Emotion as Feeling: Do We Need This for Interaction?

What is the use of felt emotions for interaction? While with simulated emotions you can have the interactive phenomenon of *contagion* (if I'm embarrassed you are embarrassed, if I'm bored or sorrowful you are bored, etc.) you cannot have *real* "empathy" and *identification*. A can neither "imagine" nor feel what B is feeling. In empathy A and B should both feel something and A's feelings are the contagion of or the identification with B's feelings.

If we believe that empathy (not simply simulated empathy) is useful in affective computing, we need agents or machines able to *feel* something.

I personally do not see cases where simulation of feelings is not enough for HC interaction. It seems to me that even conversation about what I/we feel can be based on *simulated* feelings, and as for the reactive component of the empathic reaction, this can be produced without the internal feeling. But, I might be wrong, and it might be different with robots and artificial creatures.

6.3 Intuitive Appraisal in Interaction versus Cognitive Evaluations

What the difference and the use in interaction between an intuitive, reactive appraisal of the valence of a given event or agent, and an explicit reason-based evaluation of it?

For sure affective computing requires the ability of having, showing and recognising reaction of attraction/ rejection, of liking/disliking; at an immediate, automatic, and non analytical and unaware level. This cannot be replaced by a more explicit and conceptual evaluation of an event, object or agent relative to some clear goal.

To say "I do not believe that you are good for...", or "I do not trust your ability/competence in...", is very different from showing feelings of distrust, or antipathy; and its is also different from saying "he is disagreeable", or "I do not trust him, but I don't know why" (which is a way of verbalising a mere intuitive appraisal).

Both the artificial entity and the user must be allowed to feel attraction or repulsion, or to feel trust or distrust without being able to understand or to explain precisely why - simply as a form of conditioning or association- and to express this in a verbal or non-verbal way.

However, also in this case is not clear if for interaction is necessary that also the machine *feels* this appraisal, or its simulation is enough.

On the other hand, for sure in other circumstances it is also necessary that the machine is able to make explicit its *evaluation* of something relative to a given goal and to given qualities; and that it is able to understand the direct ("this is good/useful/OK/...") or the indirect ("this is intelligent/ fast/ economic/ dangerous ...") evaluations by the user.

When Picard [44] claims that an affective computer could *understand* what its owner *likes* or *dislikes* and could adapt to the user's needs, she mixes up -for both the computer and the user- appraisal with evaluation: on the one side, we should distinguish between an explicit "understanding" of the user preferences and needs, and an affective conditioned reaction; on the other side, the user "preferences"

¹¹ Also because perhaps it is true that -at least among humans- in order to be fully believable emotions must be rather spontaneous, involuntary, non simulated and controlled

might be either true evaluations and explicit preferences, or mere affective reactions.

Both systems for dealing with valence are necessary, and both are related to emotional reaction and interaction: as we saw an evaluation can start an appraisal or a complex emotional reaction like shame or guilt or pride or fear. In particular, I believe that for several reasons we have to pass from merely implicit and affective forms of trust in the machine and in the information system, to more explicit forms of trust and delegation based on true evaluations: this will be particularly important with our personal assistants, with mediator agents on the net, with our human or electronic partners in EC, and so on.

References

- [1] Arnold, M. B. (1960). *Emotion and personality*. New York: Columbia University Press.
- [2] Bechara, A., Damasio, H., Tranel, D., Damasio, A.R., (1997) Deciding Advantageously before Knowing the Advantageous Strategy. *Science* 275: 1293-95.
- [3] Bargh, J.A. and Chartrand, T.L. (1999) The Unbearable Automaticity of Being. *American Psychologist*, 54, 462-79.
- [4] Bozinovski, S. and Bozinovska, L. (1998) Emotion and Learning: Solving Delayed Reinforcement Learning Problem Using Emotionally Reinforced Connectionist Network. In D. Canamero (ed.) Proceedings of AAAI Fall Symposium "Emotional and Intelligent: The Tangled Knot of Cognition" 1998, AAAI Press, 29-30
- [5] Breazeal, C. (1998) Early Experiments using Motivations to Regulate Human-Robot Interaction. In D. Canamero (ed.) Proceedings of AAAI Fall Symposium "Emotional and Intelligent: The Tangled Knot of Cognition" 1998, AAAI Press, 31-36
- [6] Canamero, D. (1997) Modeling Motivations and Emotions as a Basis for Intelligent Behavior. *Autonomous Agents'98*, ACM Press, 148-55.
- [7] Carbonell J (1980) Towards a process model of human personality traits. *Artificial Intelligence*, 15, 1980
- [8] Castelfranchi, C. (1995) Guaranties for Autonomy in Cognitive Agent Architecture. In M. Wooldridge and N. Jennings (Eds.) *Intelligent Agents*. Springer. LNAI 890, 1995, 56-70.
- [9] Castelfranchi, C. (1998) To believe and to feel: The case of "needs". In D. Canamero (ed.) Proceedings of AAAI Fall Symposium "Emotional and Intelligent: The Tangled Knot of Cognition" 1998, AAAI Press, 55-60
- [10] Castelfranchi, C. Modeling Social Action for AI Agents. Invited paper. In *International Joint Conference of Artificial Intelligence - IJCAI'97*, Nagoya, August 23-29 1997. pp.1567-76
- [11] Castelfranchi C., de Rosis F., Falcone R., (1997) Social Attitudes and Personalities in Agents. In *Socially Intelligent Agents, AAAI Fall Symposium Series* 1997, MIT in Cambridge, Massachusetts, November 8-10, pp.16-21.
- [12] Castelfranchi C., Falcone R., (1998) Principles of trust for MAS: cognitive anatomy, social importance, and quantification, *Proceedings of the*

- International Conference on Multi-Agent Systems (ICMAS'98)*, Paris, July, pp.72-79.
- [13] Castelfranchi, C., Falcone, R. (1999). The Dynamics of Trust: from Beliefs to Action, *Autonomous Agents '99 Workshop on "Deception, Fraud and Trust in Agent Societies"*, Seattle, USA, May 1, pp.41-54.
- [14] Castelfranchi C., Falcone R.(2000). Trust is much more than subjective probability: Mental components and sources of trust, 32nd Hawaii *International Conference on System Sciences - Track on Software Agents*, Maui, Hawaii, 5-8 January 2000.
- [15] Castelfranchi, C., Miceli, M. E Parisi, D., (1980) Invidia. *Giornale Italiano di Psicologia*, VI,1 , 1980, 95-1 19.
- [16] Castelfranchi, C e Poggi, I., (1990) Blushing as a discourse: was Darwin wrong? In R. Crozier (ed.) *Shyness and Embarrassment: Perspective from Social Psychology*, Cambridge University Press, N. Y.
- [17] Chwelos, Greg and Oatley, Keith (1994) "Appraisal, computational models, and Scherer's expert system", *Cognition \& Emotion*, 3:245-257
- [18] Conte, R. and Castelfranchi, C. *Cognitive and Social Action*. UCL Press, London, 1995.
- [19] Dahl, H. and Teller, V. (1998) In D. Canamero (ed.) Proceedings of AAAI Fall Symposium "Emotional and Intelligent: The Tangled Knot of Cognition" 1998, AAAI Press, 70-75.
- [20] Damasio, A.R. *Descartes' Error*. N.Y., Putnam's Sons, 1994
- [21] Dautenhahn, K. (Ed.) *Socially Intelligent Agents*, AAAI Fall Symposium Series 1997, MIT in Cambridge, Massachusetts, November 8-10, pp.16-21.
- [22] Dyer, Michael G. (1987) "Emotions and their computations: Three computer models", Special Issue: Cognitive science and the understanding of emotion, *Cognition \& Emotion*, 3:323-347
- [23] Elliott C (1994) Research problems in the use of a shallow Artificial Intelligence model of personality and emotions. *Proceedings of the 12° AAAI*, 1994
- [24] Elster, Jon (1993) *Sadder But Wiser? Rationality and The Emotions* London.
- [25] Frijda, N. H. (1986). *The emotions*. New York: Cambridge University Press.
- [26] Frijda, N. H. (1988). The laws of emotion. *American Psychologist*, 43, 349-358.
- [27] Frijda, Nico H. and Swagerman, Jaap (1987) "Can computers feel? Theory and design of an emotional system", *Cognition \& Emotion*, 3:235-257
- [28] Gordon, R. (1987) *The Structure of Emotions*. Cambridge, Cambridge U.P.
- [29] Haddadi, A. and Sundermeyer, K. (1996) Belief-Desire-Intention Agent Architectures. In G.M. O'Hare and N.R. Jennings (eds.) *Foundations of Distributed Artificial Intelligence*, Wiley & Sons, London.
- [30] Hayes-Roth B (1995), Agents on stage: advancing the state of the art of AI. Proceedings of *IJCAI95*
- [31] Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108, 480-498.

- [32] Lazarus, R. S., Averill, J. R., and Opton, E. M. jr. (1970). Towards a cognitive theory of emotion. In M. B. Arnold (Ed.), *Feelings and emotions* (pp. 207-232). New York: Academic Press.
- [33] Loyall A B and Bates J (1997) Personality-Rich Believable Agents That Use Language. In Proceedings of *Autonomous Agents 97*, Marina Del Rey, Cal.,106-13.
- [34] Mandler, G. (1982) The Structure of Value: Accounting for taste. In M.S. Clark and S.T. Fiske (eds.), *Affect and Cognition*. Hillsdale, NJ: Erlbaum, 3-36.
- [35] Martin, I. and Levey, A.B. (1978) Evaluative Conditioning. *Advanced Behavioural Research*, I, 57-102.
- [36] Miceli, M. and Castelfranchi, C. (1997). Basic principles of psychic suffering: A preliminary account. *Theory & Psychology*, 7, 771-800.
- [37] Miceli, M. and Castelfranchi C. (in press) The Role of Evaluation in Cognition and Social Behaviour. In K. Dautenhahn (Ed.) *Human cognition and agent technology*. Amsterdam: John Benjamins Publishing Company.
- [38] Mower, O. (1960) *Learning Theory and Behavior*, J. Wiley and Sons, N.Y.
- [39] Ortony, A. (1987) Is Guilt an Emotion? *Cognition and Emotion*, I, 1, 283-98.
- [40] Paiva, A. (2000) this volume.
- [41] Parisi, D. Motivation in artificial organisms. In G. Tascini (ed.) *Proceedings of the Workshop on Perception and Adaptive Learning*. Springer, in press.
- [42] Petta, P., Pinto_Ferreira, C., Ventura, R. (1999) Autonomy Control Software: Lessons from the Emotional (AA'99 WS on Autonomy Control Software; www.ai.univie.ac.)
- [43] Pfeifer, R. (1988) Artificial Intelligence Models of Emotion. In Hamilton, V., Bower, G. and Frijda, N. (eds.) *Cognitive Perspectives on Emotion and Motivation*, Kluwer, 287-320.
- [44] Picard, R. (1997) Does HAL cry digital tears? Emotion and computers; *HAL's Legacy: 2001's Computer as Dream and Reality*, Cambridge, 1997, 279-303
- [45] ReillyW S and Bates J, *Natural negotiation for believable agents*. School of Computer Science, Carnegie Mellon University, CMU-CS-95-164, june 1995
- [46] Scherer, Klaus R. (1984) Emotion as a multicomponent process: A model and some crosscultural data. In. P. Shaver (ed.) *Review of personality and social psychology*: Vol. 5, pp.35-63, Sage, CA.
- [47] Scherer, Klaus R. (1993) "Studying the emotion-antecedent appraisal process: An expert system approach.", *Cognition \& Emotion*, 7(3-4):325-355
- [48] Simon, H. Motivational and emotional controls of cognition. *Psychological Review*, 74, 1967, 29-39
- [49] Sloman, Aaron (1987) "Motives, mechanisms, and emotions", *Cognition \& Emotion*, 3:217-233
- [50] Sloman, A. and Croucher, M. Why robots will have emotions. In Proceedings of *IJCAI'81*, Vancouver, Canada, 1981, p. 197
- [51] Smith, C. A. and Ellsworth, P. C. (1987). Patterns of appraisal and emotion related to taking an exam. *Journal of Personality and Social Psychology*, 52, 475-488.

- [52] Staats, A. The Paradigmatic Behaviorism Theory of Emotions: Basis for Unification. *Clinical Psychology Review*, 10, 539-66, 1990.
- [53] Stolzmann, W. Learning Classifier Systems using the Cognitive Mechanism of Anticipatory Behavioral Control. In *First European Workshop on Cognitive Modelling*, Berlin, 1996, p.82-89.
- [54] Thagard, P. & Kunda, Z. (1987). Hot cognition: mechanisms of motivated inference. *Proceedings of the Annual Meeting of the Cognitive Science Society* 1987, , 753-763.
- [55] Thagard, P., (1998), Emotional Coherence: Trust, Empathy, Nationalism, Weakness of Will, Beauty, Humor, and Cognitive Therapy, Technical Report, University of Waterloo.
- [56] Velasquez, J. (1997) Modeling emotions and other motivations in synthetic agents. *Proceedings of Fourteenth AAAI*, P. 10-15.
- [57] von Cranach, M., Kalbermatten, V., Indermuhle, K., Gugler, B. (1982) *Goal-Directed Action*. London, Academic Press.
- [58] Webster, Ch. Adaptive Depression, Affective Computing, and Intelligent Processing. *Proceedings of the IEEE International Conference on Intelligent Processing Systems*, Oct.1997, Beijing, China, 1181-4.

APPENDIX - Emotion and Personality

We call *personality trait* [7] any internal state or processing mechanism of the agent that:

- differentiates it from other agents with which it is interacting or is compared;
- is relatively stable (either built in or inborn or learned, but now quite permanent) and cannot be just adopted or learned from outside on line;
- is mental;
- has to do with motivations, with the way of choosing, of reasoning, of planning and so on.

We agree with Carbonell that personalities are mainly goal based: some of them directly consist in the presence of a typical motivation or in a special importance of a given goal (ex. *sadist*, *glutton*); others can be considered as implicit goals or preferences. However, other personalities are rather based on "cognitive styles": ways of reasoning, attending, memorising, etc.

Personality Traits and Attitudes

That personality traits are stable does not mean that they are continuously relevant or active: if *x* is a *glutton*, when he is working this can be irrelevant. Some personality traits are conditional on a given circumstance: they are just temporary *attitudes*. An attitude is characterized by tests/conditions specifying the circumstance for its activation. An agent can assume an attitude or another (relatively to the same problem) depending on circumstances or partners. Therefore, we distinguish between *traits* and *attitudes*: both are constituents of personalities. An agent can change or decide about its attitude towards a given event, request, or agent, while it cannot change or decide about its personality traits: these are not subject to contextual changes or decisions.

In short: *a personality is a coherent, believable, stable, and typical cluster of traits and attitudes that are reflected in the agent's behaviour*. According to this definition, agents with different personalities must show different behaviours in similar circumstances; personalities should be coherent, by not producing contradictory behaviours in the same circumstances, and should produce "believable" behaviours. By doing this, they too are "believable".

Personality and Emotions

Emotional states are among those internal states that shape an agent's cognitive process and reaction; they can also characterise the agent. Emotion-based personalities can be defined, like *shameful*, *fearful*, *pitiful* and so on: these personalities are characterised by the agent's propensity for a given emotional reaction. However, emotions and personalities should not be mixed up with one the other, like it risks to happen in the "believable agent" domain. This is due to the fact that, in that domain, personalities are introduced just for the sake of "believability", and believability for sure requires emotional reactions [23] [30] [33] [44].

In our view:

- emotions do not necessarily imply personalities, since there might be emotional behaviours that are shared by the whole population of agents and do not characterise particular agents or individuals;
- personalities are not necessarily related to emotions: they might be just based on

- cognitive properties or styles (like a "fantasyful" agent, or a "fanatic" (commitment) agent, or an agent with an external attribution strategy and so on,
- preferences and goals,
- interactive strategies (ex. Tit-for-Tat agents; or cheaters, etc.).

Of course, it is true that these cognitive styles, and in particular preferences and goals, can make a given type of agent (or individual) exceptionally liable to some emotions. However, these emotions are not the basis for constructing and characterising that agent, though being useful to recognise it. In addition, emotions are not necessary: agents might be free from emotions while having personalities [11].

An Emotion-Based “Conscious” Software Agent Architecture

Lee McCauley¹, Stan Franklin¹, and Myles Bogner^{1,2}

Institute for Intelligent Systems and Department of Mathematical Sciences
The University of Memphis

Abstract. Evidence of the role of emotions in the action selection processes of environmentally situated agents continues to mount. This is no less true for autonomous software agents. Here we are concerned with such software agents that model a psychological theory of consciousness, global workspace theory. We briefly describe the architecture of two such agents, CMattie and IDA, and the role emotions play in each. Both agents communicate with humans in natural language, the first about seminars and the like, the second about job possibilities. IDA must also deliberate on various scenarios and negotiate with humans. In CMattie emotions occur in response to incoming stimuli from the environment and affect behavior indirectly by strengthening or weakening drives. In IDA the emotions are integrated with the “consciousness” mechanism, and bidirectionally connected with all the major parts of the architecture. Thus, emotions will affect, and be affected by, essentially all of the agent’s disparate cognitive processes. They will thus play a role in essentially all cognitive activity including perception, memory, “consciousness,” action selection, learning, and metacognition. These emotional connections will provide a common currency among the several modules of the agent architecture. These connections will also allow for the learning of complex emotions. The emotions serve to tell the agent how well it’s doing.

1 Introduction

Evidence continues to mount for the role of emotions in environmentally situated human agents [15, 7, 21]. The verdict is still out on exactly how the emotions affect cognition in humans, but there does seem to be general agreement that they play a more important role than previously believed. CMattie and IDA are environmentally situated software agents that will utilize emotional mechanisms to help direct

¹ Supported in part by ONR grant N00014-98-1-0332

² With essential contributions from the Conscious Software Research Group

attention, learn from situations that generate “good” or “bad” emotional states, and make evaluations of external stimuli and internal states. Including emotional capabilities in non-biological autonomous agents is not a new idea [3, 23, 20]. However, as more information is learned regarding the functions and architecture of emotions in humans, we must constantly reevaluate our models in light of these emerging ideas. One such idea is the notion that emotions are closely linked to almost all aspects of cognition and that they may provide the initial building blocks for conscious thought in infants [27].

Partially motivated by recent discoveries of the role emotions play in human cognition, we are reevaluating their role in the cognitive processes of our two software agents, CMattie and IDA. In CMattie emotions affect action selection by strengthening or weakening her primary motivators, her drives. They don’t directly affect perception, behaviors, the process of memory, or metacognition. What’s now known about the role of emotions in human cognition suggests that they should. Also, though emotions affect CMattie’s actions, those actions don’t affect her emotions. Not so with humans. With these issues in mind, we intend to significantly broaden the role of emotions in our more complex agent, IDA. We also intend to see that that role is bi-directional so that each of IDA’s cognitive modules also affects her emotions. In this paper we will describe the high level design for this reevaluation and reimplementation of the role of emotions in “conscious” software agents.

2 CMattie

“Conscious” Mattie (CMattie) [9] is the next incarnation of Virtual Mattie (VMattie), an intelligent clerical agent [10, 24, 28]. CMattie’s task is to prepare and distribute announcements for weekly seminars that occur throughout a semester in the Mathematical Sciences Department at the University of Memphis. She communicates with seminar organizers and announcement recipients via email in natural language, and maintains a list of email addresses for each. CMattie is completely autonomous, actively requesting information that has not been forthcoming, and deciding when to send announcements, reminders, and acknowledgements without external intervention. No format has been prescribed for any type of email message sent to her. CMattie implements a version of Hebbian type (temporal proximity) learning, and includes modules for perception (natural language understanding), action selection, metacognition, associative memory, “consciousness” and emotions. Unfortunately, her domain (seminar announcements) with respect to the emotion component may not be rich enough to require the emergence of complex emotions.

CMattie is designed to model the global workspace theory of consciousness [1, 2]. Baars’ processes correspond to what we call codelets, a name borrowed from Hofstadter and Mitchell’s Copycat system [11]. A codelet is a small piece of code capable of performing some specific action under appropriate conditions. Many codelets can be thought of as autonomous agents, making CMattie a multi-agent system in the sense of Minsky’s *Society of Minds* [17]. Almost all actions are taken at the codelet level. Her action selection mechanism chooses the next behavior, which is then implemented by lower-level codelets. These higher level behaviors correspond to

goal contexts in global workspace theory. Emotion codelets influence not only other codelets, but also indirectly influence behaviors through drives.

CMattie also has an associative memory based on a sparse distributed memory [14]. A new percept, her understanding of an incoming email message, associates with past experiences including actions and emotions. These remembered emotions, and the percept itself, activate emotion codelets that, in turn, influence current action selection. Thus, CMattie will produce actions, at least partially based on emotional content, and appropriate for the active goal context.

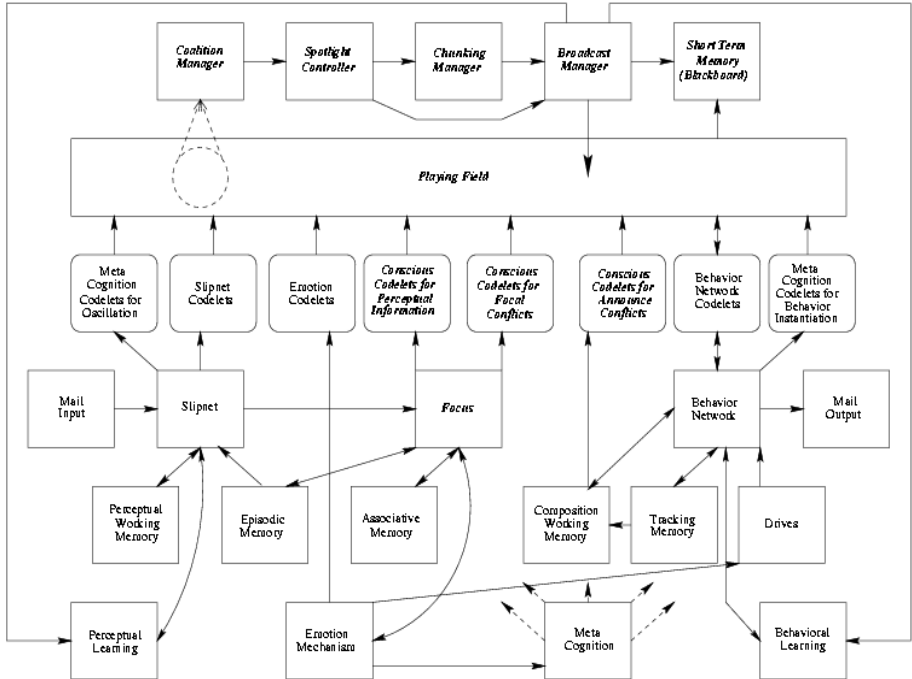


Figure 1 An overview of CMattie's architecture

3 Emotions in CMattie

In humans, emotions seem to play the role of the evaluation network. As well as affecting our choice of actions, they evaluate the results of these actions so that we may learn [21]. Emotions in an agent architecture should serve this same evaluative purpose. In CMattie [18] emotion codelets update the gain value, which is used to vary learning rates and valence, and send activation to the drives, which effects the behaviors that the system performs. The gain in CMattie and IDA is not a single value; instead, it is a vector of four real numbers that measure anger, sadness, happiness, and fear. We may later add disgust and surprise [8, 12], although, for our current purposes the current four seem to suffice. CMattie's domain is narrow

enough so that surprise and disgust would not be of great benefit. This may not be the case for IDA, who may well need them added to her repertoire.

The agent's overall emotional state at any one time is the result of a combination of the four (or six) emotions. A particular emotion may have an extremely high value as compared to the other emotions, and, consequently, dominate the agent's emotional state. For example, if a train has blocked the route to your favorite restaurant and you are hungry and in a hurry, your emotional state may be dominated by anger even though many other more subtle emotions may be active at the time. The same type of thing can occur in CMattie and IDA. In such a case the agent can be said to be angry. Do note that the agent will always have some emotional state, whether it is an easily definable one such as anger, or a less definable aggregation of emotions. No combination of emotions are preprogrammed; therefore, any recognizable complex emotions that occur will be emergent.

How exactly does CMattie determine emotional levels? The value of an individual element (emotion) in the gain can be modified when an emotion codelet fires. Emotion codelets have preconditions based on the particular state or percept the codelet is designed to recognize. For example, a message being from the system administration may be a precondition for a codelet that adds to the fear component of the emotion vector producing anxiety in the agent. How does this happen? When an emotion codelet's preconditions are met it fires, modifying the value of a global variable representing the portion of the emotion vector associated with the codelet's preconditions and may send activation to an associated drive. A two step process determines the actual value of an emotion at any one time. First, the initial intensity of the emotion codelet is adjusted to include valence, saturation, and repetition via the formula

$$a = v \frac{1}{1 + e^{\frac{(-vx+x_0)}{1.5}}} \quad (1)$$

where a is adjusted intensity at creation time, x is the initial intensity of the emotion, v is the valence $\{1, -1\}$, and x_0 is the habituation factor, which shifts the function to the left or right. The x_0 parameter models the short-term habituation of repeated emotional stimuli. Its value is increased when the same stimulus is received repeatedly within a short period of time.

During the second step in the process each emotion codelet that has fired creates an instantiation of itself with the current value for adjusted intensity and a time stamp. This new codelet will add its adjusted intensity value (not to be confused with activation) to the global variable representing its particular emotion based on the formula (modified from [20])

$$y = ae^{-b(t-t_0)} \quad (2)$$

Where a is the adjusted intensity at creation time, b is the decay rate, t the current time, and t_0 the time of creation of the codelet.

Since the emotion vector is not a single value, a single codelet will only effect one component of the vector, anger, for instance. The overall anger value for the agent would, therefore, be a summation of all of the y values for codelets that fire and that effect the anger component of the emotion vector. In this way each active emotion codelets makes its contribution to the overall emotional state. The emotional state of the agent is written to associative memory with each incoming percept. During the recall process these emotions are remembered and re-effect the emotional state of the agent by instantiating a new codelet in much the same way as an original emotion codelet would. In such a circumstance, the codelet will affect the emotional state of the agent using the previous formula adjusted for the new time of activation and with a degraded initial intensity.

There can be multiple emotion codelets, each with its own preconditions that cause it to fire. The system may fire more than one emotion codelet at a time. The resulting emotional state of the agent, represented by the gain vector, is, therefore, a combination of the recent firings of various emotion codelets. Also, multiple emotion codelets can be included in concept (chunked) codelets [13, 4], thereby learning complex emotions that are associated with a higher level concept.

So how does the emotional mechanism affect the behavior of CMattie? First, a message would come to the system and be understood by the perception module. For the purpose of this example, let us say that the message says something to the effect of, “the network will be shut down in two minutes due to a power problem.” This type of message would probably come to CMattie as a system alert, but its contents are processed in the same way as the normal email messages. After perception has categorized the message, its processed form is put into the perception registers and then moved (after some higher level processing that takes into account remembered states) to the Focus. A particular emotion codelet will then see the word “shutdown” in the Focus (it does not matter where in the Focus it is), and will increase the fear feature of the emotion vector. At the same time, this same codelet will send activation to the self-preservation drive. Several other things may be occurring simultaneously within the system, including the reading from the different types of memories and the activation of various “conscious” codelets that respond to other items in the Focus. These other codelets may spawn or activate specific behaviors in the behavior net. Along with the previous activations of behaviors in the behavior net and the newly jazzed drive, which will, through spreading activation, increase the likelihood of a self-preservation action being chosen, one behavior will get picked to fire. The probability of a self-preservation action occurring here is influenced by the amount of fear that was generated by the original emotion codelet. After a behavior is chosen and executed, there is a write back to the various memories containing all the elements of the Focus along with the state of the emotion vector and the behavior that was chosen for execution. As previously stated, this will facilitate the association of the emotional state of the system with the current state of the perceptions and the action executed. Learning can then take place at some later time that will take into account the emotions of the system as a guiding factor to how the system was doing at that time and how the associated perceptions might have affected the actions of the agent.

Another short example of how the emotions might affect CMattie’s behavior would be to note how her correspondence with a particular person could change over

time. In this case, CMattie has not received a message from a particular seminar organizer letting her know who will speak at the “Computer Science Seminar” that week and it is getting close to the time when she must send out that week’s list of seminars. She has already sent this organizer two previous reminders that week and not gotten a reply. The fact that there is missing information in the weekly seminar list, and that the time is growing short, has caused her emotional state to be high in the areas of fear (that she will not get the announcement out in time) and anger (that the seminar organizer has not sent her the necessary information). This combination of emotions might, in humans, include anxiety. The high level of anxiety might influence the template that CMattie chooses to send to this organizer. Instead of the normally “polite” version of the information request template, she might choose a more forceful version that more accurately expresses the urgency of getting the organizer’s information. Over time, CMattie could recognize that she always gets anxious because of missing information from this same organizer. Her response to this could be to choose to use the more forceful form of the information request template earlier than normal when corresponding with this organizer.

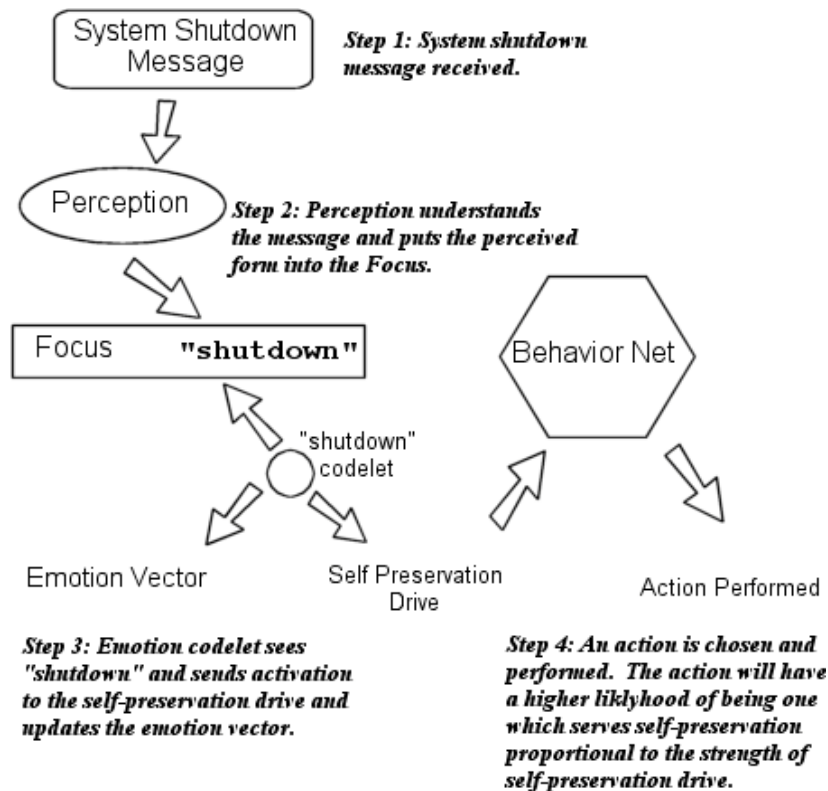


Figure 2. Example of emotions reacting to a system shutdown message

4 IDA

IDA is an Intelligent Distribution Agent for the U.S. Navy. Like CMattie, she implements global workspace theory [1, 2]. At the end of each sailor’s tour of duty, he or she is assigned to a new billet. This assignment process is called distribution. The Navy employs some 200 people, called detailers, full time to effect these new assignments. IDA’s task is to facilitate this process, by playing the role of one detailer as best she can. Designing IDA presents both communication problems and constraint satisfaction problems. She must communicate with sailors via email in natural language, understanding the content. She must access a number of databases, again understanding the content. She must see that the Navy’s needs are satisfied, for example, that the required number of sonar technicians are on a destroyer with the required types of training. She must adhere to Navy policies, for example, holding down moving costs. And, she must cater to the needs and desires of the sailor as well as is possible.

IDA will sense her world using three different sensory modalities. She will receive email messages, read database screens and sense operating system commands and messages. Each sensory mode will require at least one knowledge base and a workspace. The mechanism here will be based loosely on the Copycat Architecture [11, 29]. Each knowledge base will be a slipnet, a fluid semantic net which operates with the help of a workspace (a working memory) to allow perception (comprehension) to occur. The perception process is constructive. Each mode, other than the email and operating systems commands and messages, will understand material from a particular database, for example personnel records, a list of job openings, or a list of sailors to be assigned.

Each of IDA’s senses is an active sense, like human vision rather than human hearing. They require actions on IDA’s part before sensing can take place, for example reading email or accessing a database. One component of IDA’s action selection is an enhanced version of a behavior net [16, 19, 24]. The behavior net is a directed graph with behaviors as vertices and three different kinds of links along which activation spreads. Activation originates from internal, explicitly represented drives, from IDA’s understanding of the external world, and from internal states. The behavior whose activation is above some threshold value and is the highest among those with all preconditions satisfied becomes the next goal context as specified in global workspace theory. The several small actions typically needed to complete a behavior are performed by codelets. IDA’s behaviors are partitioned into streams, loosely corresponding to the connected components of the digraph, each in the service of one or more drives. Streams of behaviors are like plans, except that they may not be linear. Behavior streams might be interrupted during their execution or possibly not completed. Examples of IDA’s streams include Access Personnel Record, Send Acknowledgement, Offer Assignments, Produce Orders.

IDA, like CMattie, is very much a multi-agent system in the Minsky sense [17], the agents being the codelets that underlie all the higher level constructs and that ultimately perform almost all of IDA’s actions. We’ve mentioned the codelets that underlie behaviors. Others underlie slipnet nodes and perform actions necessary for constructing IDA’s understanding of an email message or of a database screen [29]. Still other codelets play a vital role in the “consciousness” mechanism.

Having gathered all relevant information, IDA must somehow select the assignments she'll offer a given sailor. Being a constraint satisfaction problem, considerable knowledge is required for making these selections. Much of this knowledge is housed in an operations research type linear functional that measures the suitability of a particular billet for a given sailor. The rest of this knowledge is found in codelets that effect the deliberation process. This process creates and evaluates a scenario to check the temporal fit of a transfer of the given sailor to a particular new billet.

IDA employs a number of different memories. The offer memory is a traditional database that keeps track of the assignments IDA has offered various sailors. For cognitive modeling purposes this memory can be considered to be external, comparable to a human keeping notes. IDA's intermediate term memory acts as an episodic memory, providing context for email messages and for the contents of database screens. It'll be implemented as a case-based memory to facilitate case-based learning. IDA's associative memory associates memories, emotions and actions with incoming percepts as well as with internal events such as deliberations. It is implemented by an extension of sparse distributed memory [14]. Some of IDA's action selection codelets act as a kind of template memory holding the various small text scripts that IDA uses to compose commands to access databases or issue orders, and to compose messages to sailors.

Global workspace theory postulates the contents of "consciousness" to be coalitions of codelets shined on by a spotlight. Imagine a codelet workspace populated by many active codelets each working, unconsciously and in parallel, on its own agenda. Coalitions of codelets that arise from novel or problematic situations seek out the spotlight. The information contained in this coalition is broadcast to all the other codelets, active or not. The idea is to recruit resources in the form of relevant codelets to help in dealing with the novel or problematic situation. It seems that in humans almost any resource may be relevant depending on the situation. Global workspace theory asserts that consciousness attacks the problem of finding the relevant resources by brute force. Broadcast to all processes. IDA uses this method.

To do so, she needs a coalition manager, a spotlight controller, a broadcast manager and "consciousness" codelets [5]. The coalition manager groups active codelets into coalitions according to the strength of the associations between them, and keeps track of them. If a collection of codelets is associated above a certain threshold level, these codelets are considered to be in a coalition. The spotlight controller determines which coalition becomes "conscious" at a particular time. It calculates the average activation level of each of the coalitions by averaging the activation levels of the coalition's codelets. The spotlight then shines on the coalition with the highest average activation level. Once the spotlight controller has determined a "conscious" coalition, it notifies the broadcast manager that is responsible for gathering information from the "conscious" coalition, and sending it to all of IDA's codelets. As prescribed by global workspace theory, messages are small and understood by only some of the agent's codelets. Specifically, the broadcast manager gathers objects labeled for broadcast from the codelets in the "conscious" coalition. These objects contain information needed for specifying the current novelty or problem. This information is then broadcast to all of IDA's codelets.

“Consciousness” codelets play a critical role in this process. A “consciousness” codelet is one whose function is to bring specific information to “consciousness.” Specific “consciousness” codelets spring into action when the information from perception is relevant to them. Some “consciousness” codelets check for conflicts among the relevant items returned from the percept and the memory. “Consciousness” codelets are designed to recognize and act on the kinds of novel or problematic situations that should be brought to “consciousness.”

5 Emotions in IDA

IDA’s emotion module, like a human’s, provides a multi-dimensional method for ascertaining how well she is doing. We will experiment with mechanisms for emotions. These may include anxiety at not understanding a message, guilt at not responding to a sailor in a timely fashion, and annoyance at an unreasonable request from a sailor. Emotions in humans and in IDA influence all decisions as to action [7].

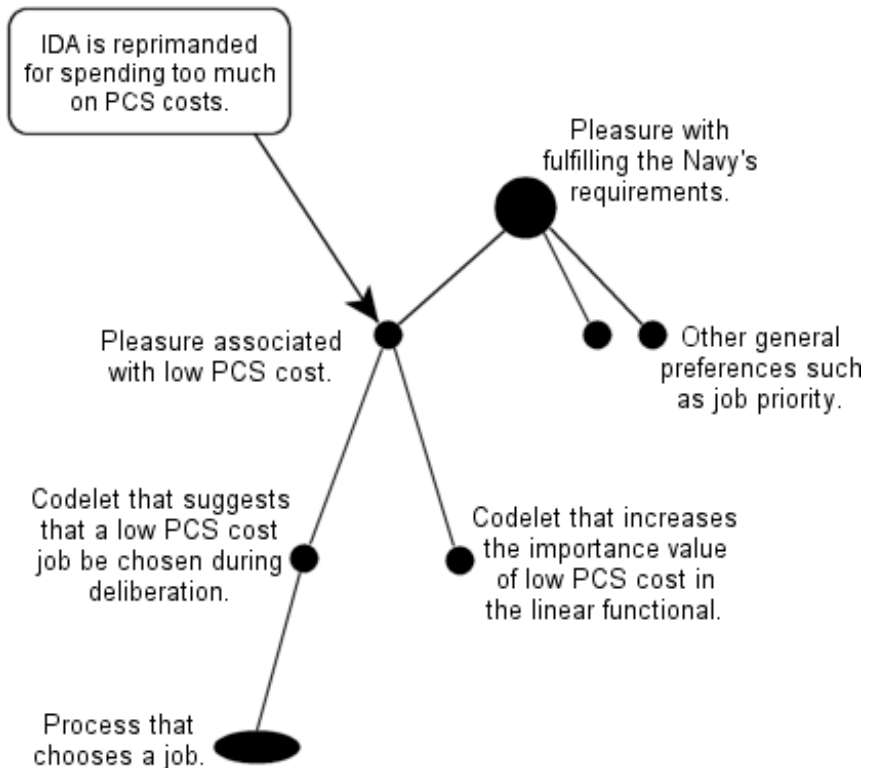


Figure 3. Illustration from example in section 5

IDA’s emotional system will need to be a good bit more robust than CMattie’s. In addition, IDA’s emotions will be more tightly integrated with her “consciousness”

mechanisms. This implementation attempts to model an emerging perspective on the relationship between emotions and the rest of cognition, in particular, the “consciousness” module. Arguments have been made that the concept of a “limbic system” that is largely separate from the rest of the brain is at best misleading. It has been suggested that the distinction between areas of the brain that are considered limbic and non-limbic cannot be made due to the incredible interconnectedness and pervasiveness of the limbic system [27]. In other words, it is not clear what aspects of cognition are being conducted with or without the aid of emotions. It seems that each time we learn a bit more about how the brain processes emotions, we are forced to reevaluate our notions of what makes emotional cognition different from all other cognitions. As Aaron Sloman has pointed out on numerous occasions, the functions that emotions seem to play can be accomplished with a complex pattern filter and alarm system [22]. CMattie is an example of just this sort of system.

What happens in humans, however, seems to be *much* more complex than what an alarm type system could produce. The first step in trying to model this complexity in IDA will be to meld portions of the emotion and “consciousness” mechanisms borrowed from CMattie. As described above, IDA’s “consciousness” module depends on codelets that each look for some particular event. These codelets, upon recognizing their preprogrammed event, activate themselves and “helper” codelets and attempt to get into the spotlight of “consciousness”. With the exception of actively trying to reach “consciousness” and recruiting other codelets, emotion codelets look very much like “consciousness” codelets, often even looking for the same event. To meld these two tasks we simply add to some of the “consciousness” codelets the ability to change the emotion vector, and link their activation to the amount of emotional change produced by that codelet.

The next step attempts to provide massive interconnectedness between the emotional mechanisms and the rest of the major cognitive areas of the system. A network is built up by connecting the “consciousness”/emotion codelets to key behaviors, goals, drives, perception codelets, etc. The links of this network are to have weights and carry activation. Weights will decay with disuse. Each use tends to decrease the decay rate. Weights will increase according to a sigmoidal function of any activation carried over the link, allowing for Hebbian style learning. The product of weight and carried activation is added to the activation already present at the head of the link. Spreading activation then becomes the common currency that integrates the separate modules that use these constructs.

As with humans, emotions in IDA provide the primary evaluative mechanism for the system. An example for how this affects IDA’s behavior is illustrated in Figure 3 above. In this situation, IDA is reprimanded because she has spent too much of the PCS (Personnel Change of Station) budget too quickly. With each new job that is assigned, a PCS cost is calculated for what it will cost the Navy to move the sailor in question from his or her current station to the new station. The Navy would like to minimize this cost. The result of the reprimand (along with unhappiness for being reprimanded) is that activation is sent to the emotion codelet that represents a general pleasure or preference for low PCS cost. Activation is sent, in turn, from this codelet through its links to two more specific emotion codelets associated with PCS cost. These two codelets will both have a direct influence over certain decisions and/or behaviors that will tend to make IDA more likely to select jobs for sailors that have

lower PCS costs (as opposed to, say, the sailors location preference). In addition, some smaller amount of activation is sent up the tree to a more general emotion codelet, in this case one which feels pleasure at meeting the Navy’s requirements. In this example, IDA would not only tend to choose jobs with lower PCS costs in the future but would also, to a lesser extent, tend to put more weight on jobs that adhere more closely to Navy policy. The example demonstrates how the emotion mechanism spreads activation of emotion codelets in both directions along its links and creates a more robust response to a situation.

This formulation provides several benefits that conform to current evidence regarding emotions. First, emotions seem to be pervasive in the human brain, linking activities that occur in relatively distant areas with a consistent value judgement. By connecting all of the major systems of IDA, we allow activities that occur in one area to effect what happens in other areas even though they seem quite disparate. Second, complex emotions can be learned over time based on their temporal cooccurrence with a situation that results in a particular emotional response. By maintaining the functionality of the emotional mechanism in CMattie, IDA will still be able to learn complex emotions through remembered emotional states and through association with other codelets that are in “consciousness” at the same time [5]. Also, the activation network as described above will allow for the learning of link weights thereby affecting the influence of emotions on actions, and vice versa. Finally, emotions can both trigger certain reactions and can be triggered by actions. In humans, for example, the physical act of singing a joyful song, even forced, can make a person temporarily feel better when they are sad. In IDA, the spreading of activation in the emotional network can occur in any direction. Thus a “consciousness”/emotion codelet can trigger an active response. In the other direction an action can trigger an emotional response or even instigate a group of codelets coming into “consciousness” that would otherwise not have been triggered.

Note that it is not the mechanisms used here to model emotions, codelets, spreading activation, etc., that are unique, but the *way* that these mechanisms are being used, and the functions that they perform within the larger system. The point is that the mechanisms involved in emotions in humans or otherwise are not “special.” The same can be said about most any other part of the human brain. The visual cortex, for instance, uses the same basic mechanisms as the temporal lobe yet play different roles within the system. It is not their substance but their organization and interconnections that set them apart. On the boundaries between areas of the brain, the naming of one neuron as being part of a module and not part of another is not only counter productive but impossible. We give names to different parts or functions of areas of the brain because it is easier to comprehend and analyze them that way, but we must keep in mind that they are all still part of the same highly parallel, highly integrated system. Emotions are much like any other cognitive function and should not be viewed as magical or special – they just perform a different function.

6 Related Work

Several other authors have proposed systems that are comparable [6, 25, 26]. The most similar of these methods is Velasquez's Cathexis, which has been used in several agent systems.

As with Cathexis, the emotion mechanism in IDA is tightly interwoven into the codelets (nodes for Velasquez) of the various systems. This interconnectedness affects the behavior of these systems differently, however, because of the nature of the implementations. In IDA, for example, we use a different mechanism for perception than we do for action selection. This means that the two modules cannot directly "talk" to each other and must pass information through the focus. The emotion mechanism provides a direct way for separate modules to affect each other's operation. For CMattie, emotions affect the behavior of the system through the activation of appropriate drives which, in turn, effect the activation level of behaviors in the behavior net.

Another difference between these systems involves the way that emotions are remembered. In IDA, the emotion vector can be remembered in associative (Sparse Distributed) and case based memory. This stores the current state of the vector along with the state of the focus and the behavior chosen for activation. It is important to note that this associates the whole vector with the state of the system at a particular time, it does not associate the change in the emotion vector with the element that triggered that change. IDA can also remember emotional activation when individual emotion triggers become associated with their effects over time via the links between emotion codelets and other codelets that tend to be active at the same time.

Both Canamero and Velasquez's work shares with our mechanism the aspect of using emotions to bias the other cognitive elements of the system. This function serves two purposes; the first being the ability to assist cognitive judgements by providing a priming for paths that lead to emotionally desirable outcomes. The second purpose is the ability to facilitate default responses to perceptions or stimuli to which the cognitive mechanisms have not yet been able to react. In both cases the emotions can be viewed as telling the agent how well it's doing.

7 Conclusion

In light of continuing developments in the field of emotion research and functional gaps in the emotional architecture for CMattie, modifications to this architecture have been described for use in IDA. It is not yet known, since IDA is not yet fully implemented, how these enhancements will ultimately affect the operation of the system. This new architecture will provide IDA with an emotional mechanism that more closely resembles that of humans. These mechanisms are more closely integrated with the "consciousness" mechanism, and connected to all of the major parts of the system in a bi-directional manner. Thus, emotions will affect, and be affected by, essentially all of the agent's disparate cognitive processes. For this reason, emotions will play a role in essentially all cognitive activity including perception, memory, "consciousness," action selection, learning, and metacognition

and will provide a common currency among the several modules of the agent architecture. These connections will also allow for the learning of complex emotions through the refinement over time of the weights on the links.

Finally, must a system have emotions for intelligence? The same question could be asked of almost any other function of the brain. Must a system have sight, particular reactive modules, or metacognition? There is probably no requirement for any one particular piece or function at this specific level. Although a system may exhibit *more* intelligence with the addition of emotions, it is probably not a necessary component. With that said, we still believe that the addition of emotions to a system can be very useful both for interacting with emotional humans and for the enhancement of cognitive abilities.

References

1. Baars, B. J. 1988. *A Cognitive Theory of Consciousness*. Cambridge: Cambridge University Press.
2. Baars, B. J. 1997. *In the Theater of Consciousness*. Oxford: Oxford University Press.
3. Bates, J. Loyall., B. A., and Reilly W. S. 1991. *Broad Agents*. In: *Proceedings of the AAAI Spring Symposium on Integrated Intelligent Architectures*. In: *SIGART Bulletin, Volume 2, Number 4, August 1992*. Stanford University.
4. Bogner, M. 1999. Realizing "consciousness" in software agents. Ph.D. Dissertation. University of Memphis.
5. Bogner, M., U. Ramamurthy, and S. Franklin. in press. "Consciousness" and Conceptual Learning in a Socially Situated Agent. In *Human Cognition and Social Agent Technology*, ed. K. Dautenhahn. Amsterdam: John Benjamins.
6. Canamero, D. 1997. Modeling Motivations and Emotions as a Basis for Intelligent Behavior. In: *Proceedings of AGENTS '97*. New York: ACM Press.
7. Damasio, A. R. 1994. *Descartes' Error*. New York: Gosset; Putnam Press.
8. Ekman, P. 1992. An Argument for Basic Emotions. In *Basic Emotions*, Ed. N. L. Stein, and K. Oatley. Hove, UK: Lawrence Erlbaum.
9. Franklin, S. 1997. Global Workspace Agents. *Journal of Consciousness Studies* 4:322–334.
10. Franklin, S., A. Graesser, B. Olde, Song H., and A. Negatu. 1996. *Virtual Mattie—an Intelligent Clerical Agent*. *AAAI Symposium on Embodied Cognition and Action, Cambridge MA.* : November.
11. Hofstadter, R. D., and Mitchell, M. 1994. *The Copycat Project: A model of mental fluidity and analogy-making*. In: *Advances in connectionist and neural computation theory, Vol. 2: Analogical connections*, Eds. K. J. Holyoak & J. A. Barnden. Norwood N.J.: Ablex.
12. Izard, C. 1993. Four Systems for Emotion Activation: Cognitive and Noncognitive Processes. *Psychological Review* 100:68–90.
13. Jackson, J. V. 1987. Idea for a Mind. *Siggart Newsletter*, 181:23–26.
14. Kanerva, P. 1988. *Sparse Distributed Memory*. Cambridge MA: The MIT Press.

15. LeDoux, J. E. and Hirst, W. 1986. *Mind and Brain: Dialogues in Cognitive Neuroscience*. Cambridge: Cambridge University Press.
16. Maes, P. 1990. How to do the right thing. *Connection Science* 1:3.
17. Minsky, M. 1986. *The Society of Mind*. New York: Simon & Schuster.
18. McCauley, T. L., and S. Franklin; 1998 An Architecture for Emotion; AAAI Fall Symposium Emotional and Intelligent: The Tangled Knot of Cognition"; AAAI; Orlando, FL.
19. Negatu, A., and S. Franklin; 1999 Behavioral learning for adaptive software agents. Intelligent Systems: ISCA 5th International Conference; International Society for Computers and Their Applications - ISCA; Denver, Colorado; June 1999.
20. Picard, R. 1997. *Affective Computing*. Cambridge MA: The MIT Press.
21. Rolls, E. T. 1999. *The Brain and Emotion*. Oxford: Oxford University Press.
22. Sloman, A. 1987. Motives Mechanisms Emotions. *Cognition and Emotion* 1:217-234.
23. Sloman, A., and Poli R... 1996. *SIM_AGENT: A toolkit for exploring agent designs in Intelligent Agents*. In: *Vol. II (ATAL-95)*, Eds. M. Wooldridge, J. Mueller & M. Tambe. : Springer-Verlag.
24. Song, H., and S. Franklin. forthcoming. A Behavior Instantiation Agent Architecture.
25. Velasquez, J. 1997. Modeling Emotions and Other Motivations in Synthetic Agents. In: *Proceedings of AAAI-97*.
26. Velásquez, J. 1998. When Robots Weep: Emotional Memories and Decision-Making. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, 70-75. Menlo Park, CA: AAAI Press.
27. Watt, D.; 1998 Emotion and Consciousness: Implications of Affective Neuroscience for Extended Reticular Thalamic Activating System Theories of Consciousness; <http://www.phil.vt.edu/ASSC/esem4.html>; On Line Web Seminar at the Association for the Scientific Study of Consciousness.
28. Zhang, Z., and S. Franklin. forthcoming. Metacognition in Software Agents Using Fuzzy Systems. .
29. Zhang, Z., S. Franklin, B. Olde, Y. Wan, and A. Graesser. 1998. Natural Language Sensing for Autonomous Agents. In *Proceedings of IEEE International Joint Symposia on Intelligence Systems* 98.

Redesigning the Agents' Decision Machinery

Luis Antunes and Helder Coelho

Faculdade de Ciências, Universidade de Lisboa
Campo Grande, 1749-016 Lisboa, Portugal
Ph: +351-1-7500087, Fax: +351-1-7500084
{xarax,hcoelho}@di.fc.ul.pt

Abstract. In a multi-agent system, agents must decide what to do and by what order. Autonomy is a key notion in such a system, since it is mainly the autonomy of the agents that makes the environment unpredictable and complex. From a user standpoint, autonomy is equally important as an ingredient that has to be used with parsimony: too much leads to agents fulfilling their own goals instead of those of the user, too little renders agents that are too dependent upon user commands and choices for their execution. Autonomy has a role in deciding which are the new goals of the agent, and it has another in choosing which of the agent's goals are going to be addressed next. We have proposed the BVG (Beliefs, Values, Goals) architecture with the idea of making decisions using multiple evaluations of a situation, taking the notion of value as central in the motivational mechanisms in the agent's mind. The agent will consider the several evaluations and decide in accordance with its goals in a rational fashion. In this paper we extend this architecture in three different directions: we consider the source of agent's goals, we enhance the decisional mechanisms to consider a wider range of situations, and we introduce emotion as a meta-level control mechanism of the decision processes.

1 Introduction

We consider a setting in which multiple agents interact in a shared environment. Usually, this environment is computer-simulated. Sometimes it is self-contained and agents are used in experiments to draw conclusions about socially relevant phenomena; in other cases, there is a user to whom the agent responds to, and a certain amount of subservience is expected from the agent.

Whichever the complexity of agents, they must possess a decision component. Even a compile-time pre-specified agent will be of little use if it is not ready for a certain extent of non-forecast possibilities. As the environment gets more demanding in terms of unpredictability (at least a priori unpredictability) more complex should our agent be in what respects to decision flexibility. The designer must have the means to specify what is expected from the agent even in a new environment s/he has never considered. With the advent of mobile computation and huge, varied artificial

environments (such as the internet), we have to enhance our agents with autonomous decision skills.

There is a strong and intertwined relation between the decision (especially, choice) and emotional mechanisms (cf. [14], and the architecture proposed in [5]). In this paper we will present an enhanced overall decision mechanism that is able to incorporate this nested relation. Our account of emotions as meta-level control influences over the decision machinery is yet preliminary: the decisions produced by our previous model are too clean, emotion-free. This paper is also the tentative answer to a challenge about how emotions influence our whole life. Our answer follows the ideas of [5]: values are in charge of filtering candidates for later decision taking; emotions control the overall decision machinery.

1.1 Decisions and Rational Agents

When confronted with a decision situation, an agent is defined as rational if he decides in such a way that pursues his self-interest. A classical way of defining self-interest is by adopting utility theory [27], that requires the agent to know in advance all possible situations and be prepared to express his preference between any two states of the world. Not only do these conditions seem difficult to be fulfilled, but also this theory leads to interesting decision paradoxes that show its limitations [17].

An attempt to escape from this kind of bounded rationality was the BDI (Belief, Desire, Intention) agent model [25]. Here, commitment to past decisions is used as a way to decrease the complexity of decisions, since committed intentions constrain the possibilities for the future, and are only abandoned when fulfilled or believed impossible to fulfil. The preferences of the agents are represented by their desires, and these will be transformed in intentions through a deliberation process.

Simon [29] proposed the idea of aspiration levels along multiple, non comparable dimensions that characterise a decision problem. Aspirations are the minimum standards some solution must meet in order to be adopted. The agent adopts and selects for execution the first solution that meets all of the aspiration levels.

1.2 Agents with Values

In a similar line of reasoning, we have addressed the issue of choice, as one of the central components in the agent's decision machinery [1, 2, 3]. We have proposed the use of multiple values to assess a decision situation. A value is a dimension against which a situation can be evaluated. By dimension we mean a non empty set endowed with an order relation. Most interesting situations from the decision standpoint will have several such dimensions, and so most decisions are based on multiple evaluations of the situation and alternative courses of action. The agent's choice machinery becomes more clear, as agents express their preferences through the use of this multiple value framework. Choice is performed by collapsing the various assessments into a choice function, that cannot be considered equivalent to a utility function, since it is computed in execution time. The multiple values framework we defend can encompass Simon's aspiration levels, but it is more general, allowing for

further flexibility, as is shown in [3]. In [2], we present an example of the use of values in a decision problem: imagine someone wants to solve some trouble with his computer. There are several ways of achieving this goal. We show how a value-endowed agent would successively try to do this, given several failure scenarios, and using as values the probability of success of a given action, the level of patience of our agent, the predicted time delay of the action to perform, and the degree of dependence from other agents each action implies.

The coexistence of these values in mind further allows the enhancement of the adaptability decision capabilities by feeding back assessments of the quality of the previous decision into the agent's decision process. Our agents' decisions no longer depend solely on the past events as known at design time. Instead, events are incorporated into the decision machinery as time passes, and the components of those processes evolve continuously to be aggregated just when a decision is needed. This is done by feeding back evaluative information about the results of the decision taken by the agent. In [2] this assessment of (the results of) the decision was done by using some measure of goodness (an abstract higher value). We also suggested other alternatives, such as the agent's own values, or the designer's values (which amounts to looking for emergent features in the agent's behaviour, that is, agents decide by using some system of values, but the designer is interested in what happens to *another* set of values, to which the agents do not have access). Even in the simplest version, the power of adaptability shown by this schema surpasses by far that of choice based on the maximisation of expected utility. It is enough to remember the demands made on the agents by utility theory: they must know *in advance* all available alternatives and preferences between any two of them [15, 17].

1.3 Overview of the Paper

In this paper we expand on the functionality of the BVG architecture by considering two important extensions. First, the decision mechanism is enhanced to cope with more complex situations. Second, the agent has to decide in absence of all the relevant evaluations. On the other hand, the source of the agent's goals is addressed in the light shed by the multiple values framework: values guide adoption and generation of goals, by providing the respective mechanisms with reasons and justifications. Finally, we tackle the global issue of control, namely how do emotions influence behaviour. We are less interested in external manifestations of emotions (e.g. facial expressions) than in the ways by which emotional life internally guides the process of decision making (in terms of [14], the so-called 'secondary emotions').

In section 2 we will address the notion of autonomy, and relate it with the agent's self-interest. In section 3 we briefly present the BVG agent architecture. In section 4, we expand on the use of values to enhance autonomy in the decision process. Values illuminate the agent's source and choice of goals, by defining the agent's character. In section 5 we readdress the issue of choice, overcoming some limitations of the previous BVG implementation. In section 6 we show how to exploit the advantages of the multiple values framework in goal adoption. Section 7 expands on emotion-driven control of the decision process. Section 8 concludes by pointing out the most important contributions.

2 Principles for Autonomy

Social autonomy is meaningful for us because we want to consider agents inserted in societies characterised by complex dynamic environments. Agents will have to take into account their past histories, as well as show initiative and reactivity, in order to make decisions in possibly difficult situations. These can include unknown information, unknown social partners and even unknown environments. Autonomy can be considered from several standpoints. Castelfranchi [11] claims that autonomy is a relational concept (mainly, a social concept: an agent is autonomous just in relation to the influence of other agents) and lists various types of autonomy, as follows:

- executive (means) autonomy: an agent is autonomous relative just to the means (instrumental sub-goals), not to the ends. It could imply some level of decision (choice among means), planning, problem-solving, etc.;
- autonomy from stimuli: the agent's behaviour should be influenced by external stimuli, but not determined or imposed by them. Behaviour has no causes, but reasons, motives.

In cognitive systems, autonomy is guaranteed by “cognitive mediation:”

- goals autonomy: the system has goals of its own, not received from the outside as commands;
- belief autonomy: an agent controls its acquisition of beliefs.

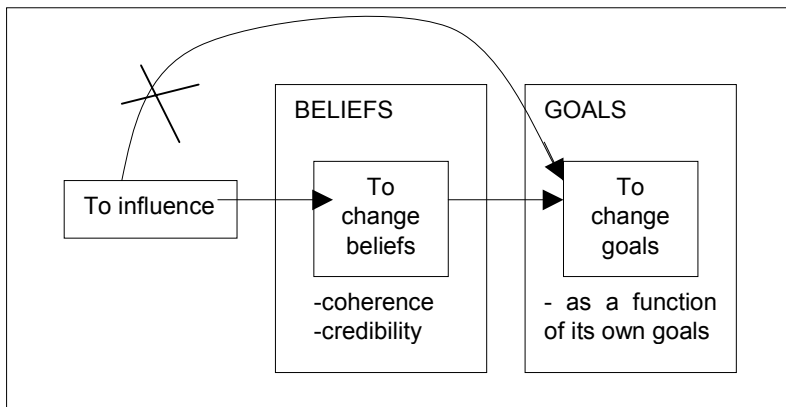


Fig. 1 Castelfranchi's Double Filter Architecture

The following postulates sketch the picture of a socially autonomous agent, and its relations, either with the outside world (including other agents), or with its own mental states, and contribute to the definition of the double filter architecture described in [11] (see fig. 1, and compare it to the one proposed in [5]). Our architecture builds on these postulates, since they solidly draw the basic picture of a self-interested agent, and one that controls its own mental states, thus providing a useful and applicable characterisation of autonomy.

- (i) it has its own goals, endogenous, not derived from another agent's will;
- (ii) it is able to make decisions concerning multiple conflicting goals (either its own goals or also goals adopted from outside);
- (iii) it adopts goals from outside, from other agents, it is liable to influencing (...);
- (iv) it adopts other agents' goals as a consequence of a choice among them and other goals (adoption is thus neither automatic nor rule-governed, and it is not simply required that the goals should not be inconsistent with the current ones);
- (v) it adopts other agents' goals only if it sees the adoption as a way of enabling itself to achieve some of its own goals (i.e. the autonomous agent is a self-interested agent);
- (vi) it is not possible to directly modify agent's goals from outside: any modification of its goals must be achieved by modifying its beliefs;
- (vii) it is impossible to change automatically the beliefs of an agent. The adoption of a belief is a special "decision" that the agent takes on the basis of many criteria and checks. This protects its cognitive autonomy.

We should notice that this version of Castelfranchi's architecture does not take emotions into account (but cf. Castelfranchi's address in [21]).

3 The BVG Architecture

While following the principles of the double filter architecture, let us draw a new agent cognitive architecture that incorporates the idea of multiple values (fig. 2), and does not include emotions. This new sketch expands on the one presented in [2], where the primary focus was laid on choice. Now, we turn our attention towards control.

We can see the execution cycle of an agent divided in three phases: perception, deliberation and execution of an appropriate action. This description is a crude simplification of a much more complex process, involving several parallel threads, with intricate relations among them. For instance, the perception is clearly guided by the agent's current focus of attention, which cannot be dissociated from its own motivations. If some relevant feature alteration is detected by the agent, it must be immediately taken into consideration, even if the decision process which included the previous value of the feature is almost concluded. We can't imagine that the world will wait until we take our decisions. But generally we have an idea about how much time we can use to decide, and how stable are the beliefs upon which we base our decisions [28].

To start with, we propose as a reference a schema of decision which includes goals, candidate actions to be chosen from, beliefs about states of the world, and values about several things, including desirability of those states. This is because we don't want to overload the architecture with too many ingredients, and related mechanisms. It is preferable to keep things manageable, and see how far we can go.

Decision is a complex, multi-staged process in an agent's mind. One stage deals with the origin of goals. Agents can either adopt goals from other agents or generate goals internally, as a result of internal processes. In another stage, goals are

considered against other objects in the agent's mind, such as beliefs (which include plans about their feasibility) and classified accordingly. For instance, we can have suspended goals, active goals, etc. Finally, among the active goals, the agent has to serialise them into execution. This is the choice phase, which we have addressed in previous papers, and will expand further herein. One way of achieving this without over-determining the behaviour of the agent, and even so reducing the deliberation time would be to have several partial ordered sets of actions, and computing the final order only when the moment arrives.

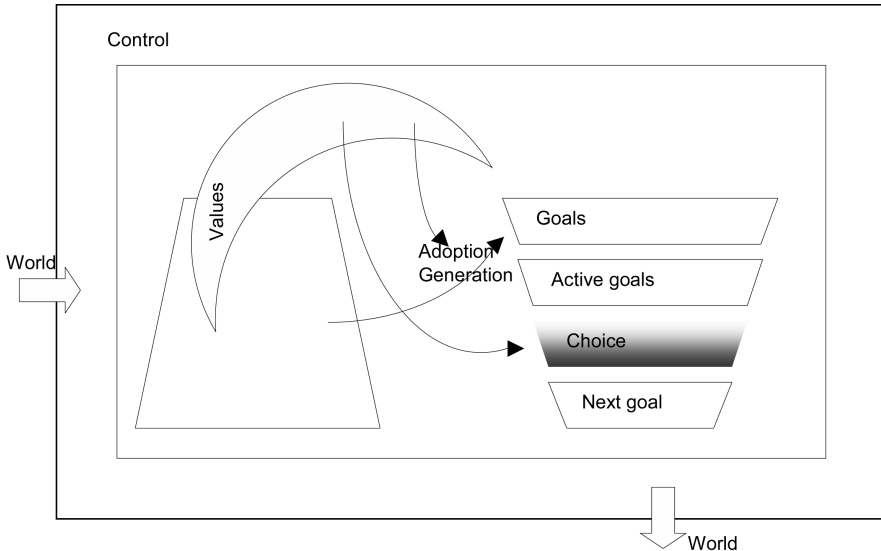


Fig. 2 The BVG architecture

The BVG architecture introduced in [2] paid special attention to choice. Now it is time to get the whole of the agent's decision machinery in context, and consider a broader picture. The role of beliefs and goals in an architecture such as the above has been thoroughly examined in the literature [13, 10]. The role of values was addressed in [1, 2], and further cleared here. The biggest challenge in BVG nowadays remains to be control. In fig. 2, control mechanisms wrap the whole of the architecture; this is meant to represent a layer responsible for a lot of important mechanisms in the agent's mind. Many of these are cognitive, along the BDI track. In [12] a set of facets were included as control mechanisms of the whole agent architecture: importance, urgency, intensity and persistence. Other control mechanisms are arguably emotional [22]. However, the association between the rational and emotional issues is still unclear [5], and our conjecture is: emotions have a dominant role in controlling the decision-making process.

As explicitly depicted in fig. 2, control mechanisms act over perception and also action. This is meant to represent the myriad of control issues involved in these two activities. Take an action that was scheduled to be executed. There are many things that can go wrong, some perhaps forecast, some completely new to the agent. Even if everything goes well, control cannot rest, for there can be opportunities to be taken,

records to be made, evaluations to be performed about how well were expectations met. What kind of machinery should this control layer possess in order to decide *only* if the agent's reasoning should or not be interrupted, and the issue of the current action be reconsidered? How many things are there to control, how many features are there to be examined to solve such a (meta-reasoning) problem [26, 15]? A lot of other roles of control are left implicit in fig. 2. Just as over external actions, control must be exerted over internal actions, such as deductions, constrainings, updates, accesses, etc. One possible solution was suggested in [31]: redundancy. We defend that agents should perform redundant calculations and have triggers that allow normal reasoning to be interrupted. Look below (section 4.2) for a possible such mechanism. [9] presents similar ideas concerning control, when the authors defend the multidimensional mind (the unidimensional mind covers only the use of utility functions) and postulate the need for a balance between action interests and high level normative motives. Imagine another situation, put forward by LeDoux [18], where a bomb is involved and people around get into panic with terrible fear. What influence did this fear have in the people's decision making? Before we further address the answer to this question let us tackle some more architectural issues.

4 Autonomy and Character

As we have seen in Castelfranchi's analysis, autonomy is involved in all of the stages of decision. An agent that generates its own goals is certainly more autonomous than another that doesn't. A further flavour of autonomy is the ability to adopt goals from other agents.

The selection of the next action to be executed results from successively constraining the set of candidates for goals. Following the path in fig. 2, we see that from the set of possible goals we extract the actual goals, either by adoption or by generation. Among these, some are considered active. For instance a goal known to be impossible to fulfil could not be considered active. Some of the active goals are then selected to be executed, and so are sorted according to some criteria.

So we have three processes by which our agent constrains its set of goals. The (1) *creation* of goals is the first phase, including both adoption and generation of goals. After that, the (2) *selection* of active goals occurs. From the active goals the agent (3) *chooses* the goals to be executed. Most of the literature on intelligent agents only focus on the selection phase, because it is the 'cleanest' one, the most technical one. It is possible to accomplish an interesting performance level relying only on technical reasons to limit the set of goals, such as a deductive machinery with *modus ponens*. See, for instance, the BDI (Belief, Desire and Intention) architecture [24, 25]. Usually the choice phase is either not addressed or implemented with very simple heuristic solutions, such as using simple utilities. Also, the creation phase, many times, is oversimplified by just postulating that the only source of goals is the agent's user.

However, when we want to define autonomous agents, we must address all of these phases, which raises the issue of the agent's character. The character can be defined as the set of collective qualities, especially mental and moral, that distinguish a person or entity. Autonomy implies difference, i. e., the agents should be free to decide

differently from one another even in the same situation. Therefore, agents can have different characters. Either in the creation (1) or in the choice (3) phases, the reasons can vary from agent to agent. In the BVG architecture, it is in these two phases that we can locate the personality traits that define the character. In the creation phase, some agents will adopt goals whereas others don't, some agents will generate goals in one way, while others may not generate goals at all, and only obey goals from their creator. In the choice phase, we look at different rationalities in different agents. Some choice is not necessarily irrational just because we cannot identify the reasons that led to it [15]. Even considering agents endowed with the same set of goals and beliefs, their different sets of values should be enough to produce differentiated behaviours. For example, a soccer robot can decide to shoot weakly at goal because he is too close to the goalkeeper and thinks there's some risk of endangering him. Finally, the influence emotions exert in decision making, can also be tuned up differently in different characters. Some agents will be bolder, or more timid, or more risk averse, etc. Subsequently (cf. section 7), we will see how this can be achieved in BVG.

4.1 Generation and Adoption of Goals

The source of goals is usually considered to be the designer of the agent, and acquisition of goals is done once and for all in an a priori fashion. This is clearly very limitative in what concerns autonomy. Acquisition mechanisms are necessary so that the agents can profit from all possibilities the environment can offer.

Castelfranchi proposes to only adopt goals when these are instrumental to other previously existing goals. This is an oversimplistic solution to the adoption problem, that is founded in architectural limitations, the fact that only beliefs and goals are available. This mechanism of adoption would lead to too rigid a structure of goals: if all goals are instrumental to previously existing goals, all sub-goals must be justified by these other higher ranked goals. Who provided these ones? Since all adoption is guided by them, they must have been adopted in design time. But then these higher (or end-) goals are the only ones important, all others could be abandoned as easily as they were adopted, if the agent discovers their uselessness for fulfilling the end-goals. So the agent's autonomy is severely restricted to executive autonomy. How can we then represent the agent's self-interest?

We propose to use values to help define the source of goals, guiding either adoption or generation of goals. When an agent identifies an object of the world as a candidate goal, it can decide to adopt it for a number of reasons. Of course if this new goal is instrumental to a previous existing goal, it should be a stronger candidate, but don't the characteristics of the existing goal matter? We think that the candidate goal should possess some added value itself. On top of the technical reasons (e.g. the new sub-goal allows the super-goal to be fulfilled), other reasons can lead to adoption. If some behaviour is perceived as leading to a situation assessed as desirable, the agent can adopt as their own the goals that triggered that behaviour. And this desirability can and should be expressed in terms of the agent's system of values.

Candidate goals come from everywhere in the agent's life. Virtually any interaction with the environment and its inhabitants can provide candidates for goals.

Mechanisms for adoption include imitation of other agents, instincts, training and learning, curiosity, intuition, etc. We can use the values framework to fine-tune these mechanisms, and even apply the same adjustment methods we used in the choice mechanism to enhance them. In section 6 we will see how these mechanisms can be defined in the BVG architecture.

Goal generation mechanisms can also be based on the agent's values. When a particular state of the world is evaluated as desirable, the agent can transform this description of the world into a goal to be achieved. Goal generation is in a lot of ways a more complex problem than goal adoption, since it has to do with creativity, and will be left out in the rest of this study.

4.2 Enhancing the Choice Mechanisms

In [2] we have seen how choice can be guided by values, and how this choice mechanism can be adapted in accordance with evaluation of successive results. But in that paper, we worked out our models by considering severe limitations that must be overridden. We had a fixed set of values, against which all candidate actions could be compared. We also had a fixed decision function (linear combination of the multiple evaluation of the candidates) and a fixed value adaptation function. Finally, we had only two types of values: aspiration-type values (e.g. probability of success), and capital-type values (e.g. time delay).

Even when considering only the single-agent case, every one of these limitations is unacceptable since they radically restrain the expressive power of the model. When the multiple agent case is considered, they are even more damaging. Consider the ability to communicate values, and consequently goals and actions. Several steps must be performed in planning the experiments, making this path follow the agent's classification according to the use of values. First, increase the number of values present in the choice setting. Then we can consider options that are characterised by values which were not necessarily foretold. That is, the agents must perform choice even in the presence of incomplete information. Afterwards, we will consider a variable set of values. Given the previous setting, this should not raise any problems, but in this case it is important to observe the agent's behaviour in the long run. Remember that value adaptation is performed over the candidate actions, even if this wasn't the only option we could have made. So we must increase the number of candidate actions, to see how the model copes, and check whether an alternative formulation should be used.

An interesting alternative could be the use of value accumulators. By adapting the idea of [6] for memory access, we can have an indicator of how much a value was referred to by other mental objects. A competition among values is cast, allowing the agent to have a relative idea of the importance of the various values. This schema does not necessarily substitute the usual schema of evaluation/choice/adaptation, it just allows the agent to subvert this in special occasions. For instance, if an option is being considered that refers to an important quantity of the highest rated value, execution stops and all calculations are restarted. This could be a simple way of implementing a pre-emptive mechanism, which we deem extremely important in an open environment.

4.3 Ontology of Values

The BVG architecture is sufficiently general to cope with a series of diversified situations. By selectively choosing the set of relevant values and associated mechanisms for choice and for value adaptation, the designer can create instances of BVG agents which are adequate to his own setting. However, we can conceive specialisations of the most general BVG agent that provide a default basis for the development of agents. One such step has already been taken [2] when we considered the notion of goodness of a decision to help recalibrate the value system that led to it. This goodness is a kind of higher value that would be present in most agent designs (see section 7).

Another strong candidate to be present in a lot of agent designs would be survival (or vital energy) as a fundamental value that would govern most decisions. If we launch a group of robots to explore Mars, we would like them to keep functioning as long as they possibly can, even if they would sometimes undermine other valued possibilities. Survival is particularly important since it is the fact that the agent keeps working that allows the result of its efforts to be exploited, passed along to its user, etc. Several years ago Sloman [30] proposed to characterise goals with three features: urgency, intensity and persistence. These are also candidates to be usually present as values when tackling a choice situation.

As with the credibility of a belief, also the probability of success of a given action is likely to be considered when choosing that action. With time, a designer can develop an ontology of values that characterise most of the choice situations his agents face, or at least to have several ontologies adapted to classes of problems.

5 New Mechanisms for Choice

In [2], our agent had a goal to fulfil that was characterised by some values. The candidate actions were all comparable according to the same values, so choice was straightforward. The agent just computed a real function of those values for all the alternatives, and the highest scorer was chosen. Afterwards, the agent looked at the results of his action by assessing the resulting state of the world against a dimension of goodness (cf. section 1 and [3]), and updated the values appropriately.

We now propose to tackle choice situations where the agent doesn't have all the relevant information. Imagine some goal G is characterised by targets for three (out of five) values: $V_1=\omega_1$, $V_2=\omega_2$, $V_3=\omega_3$. Let the candidate actions be represented by sub-goals G_1 and G_2 , with associated values, respectively: $V_1=\omega_{11}$, $V_4=\omega_{14}$, and $V_1=\omega_{21}$, $V_2=\omega_{22}$, $V_5=\omega_{51}$.

As long as choice is performed by using a linear combination of some function of the paired (e.g. ω_1, ω_{11}) values, like in [2], one can just omit the values outside of the intersection of the goal and the candidate characterisation, thus rendering the other values not redundant. If other types of choice functions are used, one must proceed with more care. Anyway, even in the simple case above, there are open problems to be dealt with. First of all, it is necessary to characterise the new adopted goal, say G_2 . Should G_2 include values for V_3 ? Should it keep values for V_5 ? We think that the

answer to both questions is positive: V_3 should be kept (with target ω_3) because we are adopting G_2 just because of G . So it is only fair that we keep whatever guides the attempts at achieving G , and those are the values V_1 , V_2 , and V_3 . For an analogous reason we should include V_5 . It could be the case that V_5 represents important evaluative notions to be considered during the attempts at G_2 , and so we mustn't give up V_5 for our future execution. In both cases, these values will help control the execution of the agent towards his goals, possibly allowing for revisions and recalculations if the chosen goals no longer serve the relevant values at stake.

6 Goal Adoption

In this section, we illustrate our ideas about goal adoption by proposing a concrete mechanism for adoption: imitation. To simplify, assume an agent perceives as a possible goal [23] some action he observed another agent carrying out (or possibly as a result of some interaction). In the multiple values framework, any object carries with it some characterisation in terms of values. Let us further assume this set is non-empty.

If we take Castelfranchi's rule for adoption, our agent will adopt this possible goal as a goal of his own, only if he perceives this new goal as serving one of his previously existing goals:

$$\text{Adopt}(\text{agentA}, \text{goal}(\text{agentB}, G_0)) \text{ iff} \\ \exists \text{goal}(\text{agentA}, G_1): \exists \text{plan}(\text{agentA}, G_1, P_1, \dots, P_n): G_0 \supset G_1$$

These are what we have called technical reasons for adoption. In BVG, the agent has reasons for adoption that are founded in his values, that represent his preferences. If we want to maintain Castelfranchi's rule, our rule of adoption by imitation could be to adopt the candidate goal if there is already a goal that shares some value with the new goal to adopt:

$$\text{Adopt}(\text{agentA}, \text{goal}(\text{agentB}, G_0, V_1=\omega_1, \dots, V_k=\omega_k)) \text{ iff} \\ \exists i \in \{1, \dots, k\}: \exists \text{goal}(\text{agentA}, G_1, \dots, V_i=\omega'_i, \dots): \omega_i * \omega'_i > 0$$

In the absence of a goal to be served by the goal candidate for adoption, we could propose another rule of adoption by imitation that would base adoption upon the values concerning the imitated agent:

$$\text{Adopt}(\text{agentA}, \text{goal}(\text{agentB}, G_0, V_1=\omega_1, \dots, V_k=\omega_k)) \text{ iff} \\ \exists \text{bel}(\text{agentA}, \text{val}(\text{agentB}, V_i=\xi_i, \dots, V_j=\xi_j)): \\ \exists \text{val}(\text{agentA}, V_i=\xi'_i, \dots, V_j=\xi'_j): \forall l \in \{i, \dots, j\} \xi_l * \xi'_l > 0$$

This mechanism (imitation) is also interesting for an emotional machinery (cf. [18]). Young children imitate emotional reactions much before they internalise the real emotions. Damasio's secondary emotions are cognitively generated, and arise later, as a result of identified "systematic connections (...) between primary emotions and categories of objects and situations" [22]. Children build these mental connections by going through imitation and game-playing: they emulate the emotional behaviour. In the panic situation mentioned above, people use imitation as an emotional-driven goal adoption strategy. There is time to consider only one or two alternative courses of action (run, dive), and then a decision must be reached.

7 Affective Reasoning

In [2], the adaptation of the agent system of values has three possible sources of evaluative information, that assess the quality of the decision taken: (i) some measure of goodness; (ii) the values that led to the decision themselves; and (iii) the designer's values. We now propose an alternative meaning for these assessment measures.

We suggest that the goodness of a decision (i) amounts to the agent's affective appraisal of the outcome produced by that decision. That is, goodness is not an abstract notion, independent of whoever performs that assessment, it is subjective, and it involves the agent that performs the assessment, in the exact conditions the assessment is made, and considering the conditions in which the decision was taken.

Alternative (ii) is the 'clean' option, for is performed in the same terms as the decision. That is, if one is already looking to optimise some things, just go and measure them. Observation subjectivity apart, the agent should have no difficulty with this. So, these values do not appear very interesting as a candidates for emotion arousal. Since goodness (i) was not considered in deciding (unless decision was taken using simple utility), the natural extension is to consider a different set of evaluative dimensions to assess the outcome of the decision (iii).

The interesting possibility about the designer's values is to consider this set to be the set of emotional features raised by the issue at stake (see figure 3d/e). These can be either inspired by the agent's own affective perception, or driven by some interaction language that interfaces to the user, or other agents. We humans frequently evaluate our decisions by examining the emotional reactions they raised in the surrounding people [18]. In fact it is often like this we notice we goofed up. Of course, we raise here a regress problem: if we want to enhance our decision capabilities by using others' views about the result of our actions, we should use our decision skills to judge whether to trust our impressions about those views (since we don't have direct access to them). To solve this problem we must accept that subjectivity has its limits: we only know what we know, and never what others know [4].

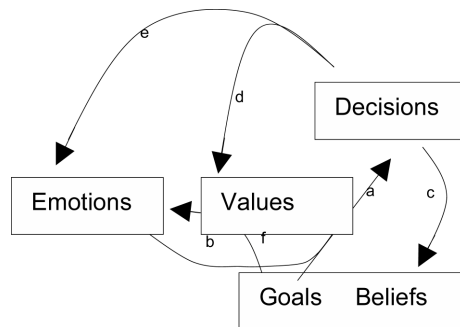


Fig. 3 The dynamics of emotions. (a) value-informed decision making; (b) emotions exert control over decision; (c) watching consequences of decisions; (d) feeding back information to enhance decision machinery; (e) getting emotional information about the results of decisions; (f) cognitive evaluations arising emotions

The view just described concerns primarily the observation of emotions raised as a consequence of the agent's behaviour (figure 3e). In a sense, it is an external view of

emotions, although considering the consequences of the emotional assessment in the calibration of the decision machinery. What is lacking here is the direct influence of emotions in decision making (figure 3b), as suggested by recent literature (especially [5]). At this point of our research, we can only point out a broad picture, which we deem coherent, and also consistent with the most recent neuroscience views on emotion and its strong influence on decision making.

In [7,8], emotions are considered as meta-level controls that help the agent organism to regulate itself in order to improve its adaptation to the environment. They have an impact on cognitive processes such as the allocation of cognitive resources, attention focus and adaptive learning, either boosting some ones or freezing others. Also here, the key idea is that emotions establish control over the decision making process. They serve as signals, or somatic markers, to be used in case retrieval, like a sort of pre-cognitive assessment of situations, based on similar experiences from the past. Emotions also determine the amount of time the agent will have for the decision process. One strategy could be to lead the agent to a quick reaction, as a way of gaining some time to spend on a more thorough decision process. For instance, in LeDoux's bomb example, one dives into the ground, and freezes any decision while considering other options.

Let us go a little further into the detail of these mechanisms. An agent should be ready for action taking at any point of its life (or it won't survive long in a hostile environment). This means that although deliberation is certainly useful and many times unavoidable, it should possess anytime character. If we take this idea to its limits, we must include in our architecture some immediate feedback mechanism, what is usually called a reactive component. We propose that this spectrum from reactive action (deliberation time equals zero) to fully deliberative action (unlimited deliberation time) can be achieved through a cumulative deliberation framework. Emotions tune up the amount of time available for the choice function to perform (for instance an additive function). In a strongly emotional stress situation (as in a state of fear, or panic), actions are taken from a small repertoire of readily accessible candidates, possibly with no deliberation at all (e.g. we do the first thing that comes to mind [18]). With some more time available, our cumulative assessment can begin: we pick a larger repertoire of agents, at the same time start the calculation of their choice value. This is done by considering the most important value, and performing the choice calculation for each of the candidates with respect to this value. At each moment, we have an assessment (a rating) that allows us to perform choice. With infinite time available, choice will be perfect, i.e. in accordance to our beliefs and evaluations about the relevant issues at stake. This broad picture is compatible with Damasio's evidence that "Normal people choose advantageously before realising which strategy works best" (the idea of importance, of valued and weighted decision), and that "non-conscious biases guide behaviour before conscious knowledge does" (emotional reactions are triggered before any rational deliberation) [5].

When we evaluate some situation, weighting pros and cons, we mix emotions and feelings with reasons, and if we succeed, we may say that the whole process is a mixture of reasons and emotions. Emotions are not simple reflexes, but patterns (collections of answers) supporting fast reactions (in order to maintain us alive), faster than cognitive actions, because from an evolution point of view the emotional machinery is older than the cognitive machinery. Quoting from [5]: "The sensory

representation of a situation that requires decision leads to two largely parallel but interacting chains of events”: emotional ones based upon previous individual experience (complex process of non-conscious signalling) and rational ones based upon processes of cognitive evaluation and reasoning.

We now take the theory of emotions in [20], in what respects the variables that affect emotion intensity. [19] lists the variables affecting anger: the degree of judged blameworthiness, the degree of deviation from personal or role-based expectations, and the degree to which the event is undesirable. We have already seen how emotions influence decision making (fig. 3b). Emotions also influence cognitive acquisition (for instance, through determining attention focus [6], and so through a decision-mediated process, see fig. 3b/a/c). Now we see what influences emotions. [19] tells us that “emotions are dependent on beliefs,” and that “emotions are positively or negatively valenced reactions.” Notice that all those prototypical variables above are cognitive assessments, evaluations about the reciprocal influences between the event and the agent’s informational and pro-active states. That is, we have beliefs, goals and (what we call) values influencing emotions. The circle is completed, and we have now cognitive values as the basis of emotion arousal (figure 3f).

8 Concluding Remarks

In this paper, we have restated the fundamentals of the BVG architecture, namely the use of multiple values to inform choice, and the feedback of assessment information to recalibrate the choice machinery. We took Castelfranchi’s notions about agent autonomy to enhance the BVG architecture in two ways. First, we considered and overcame some limitations of the previous implementation, such as the inability to deal with incomplete evaluative information. Second, we proposed to use the agent’s system of values to inform the mechanisms of creation of goals. As an example, we presented rules for adoption that could implement the mechanism of imitation. The primary concern throughout the paper is agent autonomy, and its relations with the agent’s character, or personality. We defend that the multiple values framework is especially adequate to provide the tools to successfully tackle these issues. This was further elaborated by the suggestion that not only cognitive but also emotive aspects of the agent’s reasoning can be addressed in this framework. Emotions play a determinant role in decision processing. They influence and are influenced by decisions and their consequences. Together with values, emotions provide the fundamental mechanisms for efficient and adaptable decision processing.

We could only hint at applications of these ideas. Some experiments were made, but the results are not yet conclusive. So, future research will aim at experimental demonstration of the ideas presented herein. Other research issues remain to be the increase of the agent’s adaptability by enhancing the use of the fed back information, and the expansion of an agent’s system of values as a result of interaction with other agents.

Acknowledgements

We wish to express our thanks to João Balsa, José Castro Caldas, João Faria, José Cascalho, Cristiano Castelfranchi, Rosaria Conte, Maria Miceli, Luis Moniz, Leonel Nóbrega, Pedro Rodrigues, Jorge Louçã, the editor of this volume, and the anonymous referees. This research has been carried out within the research unit LabMAC, and partially supported by project Praxis 2/2.1/TIT/1662/95 (SARA).

References

- [1] Antunes, L., Towards a model for value-based motivated agents, in proceedings of MASTA'97 (EPIA '97 workshop on Multi-agent Systems: Theory and Applications), Coimbra, October 1997.
- [2] Antunes, L. and Coelho, H., Decisions based upon multiple values: the BVG agent architecture, in Barahona, P. and Alferes, J. (eds.), Progress in Artificial Intelligence, proceedings of EPIA'99, Springer-Verlag, Lecture Notes in AI no. 1695, September 1999.
- [3] Antunes, L. and Coelho, H., Rationalising decisions using multiple values, in proceedings of the European Conference on Cognitive Science, Siena, October 1999.
- [4] Antunes, L., Moniz, L. and Azevedo, C., RB+: The dynamic estimation of the opponent's strength, in proceedings of the AISB Conference, IOS Press, Birmingham 1993.
- [5] Bechara, A., Damasio, H., Tranel, D. and Damasio, A. R., Deciding advantageously before knowing the advantageous strategy, Science, Vol. 275, 28 February, 1997.
- [6] Botelho, L. and Coelho, H., Emotion-based attention shift in autonomous agents, in Müller, J.P., Wooldridge, M.J., Jennings, N.R., Intelligent Agents III, Agent Theories, Architectures, and Languages, ECAI'96 Workshop (ATAL), Springer-Verlag, Lecture Notes in AI no. 1193, 1997.
- [7] Botelho, L. and Coelho, H. Adaptive agents: emotion learning, Proceedings of the Workshop on Grounding Emotions in Adaptive Systems, Fifth International Conference of the Society for Adaptive Behaviour 98 (SAB'98), Zurich, August 21, pp. 19-24, 1998.
- [8] Botelho, L. and Coelho, H. Artificial autonomous agents with artificial emotions, Proceedings of the 2nd International Conference on Autonomous Agents (Agents'98), Minneapolis/St. Paul, May 10-13, pp. 449-450, 1998.
- [9] Caldas, J. M. C. and Coelho, H. The origin of institutions, socio-economic processes, choice, norms and conventions, Journal of Artificial Societies and Social Simulation (JASSS), Vol. 2, No. 2, 1999.
- [10] Castelfranchi, C., Social Power. A point missed in Multi-Agent, DAI and HCI., in Decentralized AI - Proceedings of MAAMAW'90, Demazeau, Y. and Müller, J. P. (Eds.), Elsevier Science Publishers B. V., Amsterdam, 1990.

- [11] Castelfranchi, C., Guarantees for autonomy in cognitive agent architecture, in Wooldridge, M.J., Jennings, N.R., *Intelligent Agents, Agent Theories, Architectures, and Languages*, ECAI'94 Workshop (ATAL), Springer-Verlag, Lecture Notes in AI no. 890, 1995.
- [12] Corrêa, M. and Coelho, H. From mental states and architectures to agents' programming, in proceedings of the 6th Ibero-american Conference in Artificial Intelligence, Lisboa, October 5-9, 1998, Coelho, H. (ed.), "Progress in Artificial Intelligence - Iberamia'98," Lecture Notes in AI no. 1484, Springer-Verlag, pp. 64-75, 1998.
- [13] Cohen, P. R. and Levesque, H. J., *Intention=Choice+Commitment*, in proceedings of AAAI'87, 1987.
- [14] Damasio, A., *Descartes' Error: Emotion, Reason, and the Human's Brain*, G. P. Putnam's Sons, New York, 1994.
- [15] Doyle, J., *Rationality and its roles in reasoning*, Computational Intelligence, vol. 8, no. 2, May 1992.
- [16] Elliott, C., *Research problems in the use of a shallow artificial intelligence model of personality and emotion*, in proceedings of AAAI'94, 1994.
- [17] Hollis, M., *The Philosophy of Social Science - An Introduction*. Cambridge: Cambridge University Press, 1994.
- [18] LeDoux, J., *The Emotional Brain*, Touchstone (Simon and Schuster), New York, 1998.
- [19] O'Rorke, P. and Ortony, A., *Explaining Emotions (Revised)*, Tech. Rep. ICS-TR-92-22, University of California, Irvine, Department of Information and Computer Science, June 1993.
- [20] Ortony, A., Clore, G. and Collins, A., *The Cognitive Structure of Emotions*, Cambridge University Press, Cambridge MA, 1988.
- [21] Paiva, A. and Martinho, C., *Proceedings of the workshop on Affect in Interactions (towards a new generations of interfaces)*, of the 3rd i3 annual conference, Siena, October 1999.
- [22] Picard, R. W., *Affective Computing*, M.I.T. Media Laboratory Perceptual Computing Section Technical Report No. 321, November 1995.
- [23] Pollack, M. E., *Overloading Intentions for Efficient Practical Reasoning*, Noûs, vol. 25, no. 4, 1991.
- [24] Pollack, M. E. and Ringuette, M., *Introducing the tileworld: experimentally evaluating agent architectures*, in proceedings of AAAI'90, 1990.
- [25] Rao, A. S. and Georgeff, M. P., *Modeling Rational Agents within a BDI-Architecture*, in proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning, Morgan Kaufmann, 1991.
- [26] Russell, S., *Rationality and Intelligence*, Artificial Intelligence, vol. 94 (1-2), Elsevier, July 1997.
- [27] Russell, S. and Norvig, P., *Artificial Intelligence: A Modern Approach*, Prentice Hall, 1995.
- [28] Russell, S. and Wefald, E., *Do the right thing - studies in limited rationality*, The MIT Press, 1991.
- [29] Simon, H., *The Sciences of the Artificial* (3rd edition), the MIT Press, Cambridge, 1996.

- [30] Sloman, A., Motives, Mechanisms and Emotions, in *Emotion and Cognition* 1, 3, 1987.
- [31] Sloman, A., Prolegomena to a Theory of Communication and Affect, in Ortony, A., Slack, J., and Stock, O. (Eds.), *AI and Cognitive Science: Perspectives on Communication*, Springer-Verlag, 1991.

Artificial Emotion and Emotion Learning: Emotions as Value Judgements

Stevo Bozinovski

RoboCup Team, Behavior Engineering Group, Institute for Autonomous Intelligent Systems,
German National Center for Information Technology, D-53754 Sankt Augustin, Germany
on leave from

Laboratory for Intelligent Machines, Bioinformation Systems and System Software
Electrical Engineering Department, University of Skopje, MK-1000 Skopje, Macedonia
bozinovs@rea.etf.ukim.edu.mk

Abstract. The paper discusses the role of emotion in learning, presenting a connectionist learning architecture that indeed uses the concept of emotion in its learning rule. Presenting a working learning architecture based on the concept of emotion as value judgement, the paper contributes toward the efforts in answering the long-lasting question stated in psychology and philosophy: what is an emotion. The description of the architecture is followed by description of an experiment, in which the architecture is used as a learning controller in a process of training a simulated goalkeeper robot for a RoboCup match.

1 Introduction

Emotion has always been recognized as an important feature underlying human behavior and has been considered as a distinct research area in psychology and philosophy. But, the concept of emotion, being natural in humans, is rather difficult to implement in artificial agents. In artificial agents research, although most of the effort has been put on cognitive part of artificial beings, (under the name of Artificial Intelligence AI), in the last twenty years there have been several pointers toward the issue of feelings and emotions (Artificial Emotion, AE) as needed features for developing an artificial intelligent creature (e.g. [32], [5], [23]). More recently, a critical mass of researchers in the area induced dedicated meetings to present the obtained results (see [13], [14], [36], [25]). Several issues in the emotional agent design emerged, among others the problem of *emotion-based learning* in autonomous agents. The issue has been addressed from several viewpoints [7], [17], [35]. In essence, the problem is how to define emotion-based learning in a robot, and how to implement it in a useful and straightforward manner.

This paper deals with philosophical issues and a computational model of emotion that can be used in the learning system of a robot. We will start our discussion with appreciation of some philosophical and psychological findings: that the emotion can be considered as a process of internal value judgement. After that we will give a taxonomy of learning paradigms in which we place the *emotion learning* paradigm. Then we will describe an agent architecture which actually implements a computational model of emotion learning. We will present experimental evidence of an application of such an agent architecture as a controller of a robot that can be

trained to become a goalkeeper of a robot football team in a RoboCup match. At the end of the paper we will give a comment regarding the related work.

2 A Lesson from Psychology: Emotion as Internal Value Judgement

Among various taxonomies of emotion in philosophy and psychology, here we will briefly consider the taxonomy proposed by Calhoun and Solomon [12]. They divide emotional theories into sensation, physiological, behavioral, evaluative, and cognitive. Example of a theory of emotion saying essentially that emotion is a physiological reaction, essentially its familiar sensory accompaniment - a "feeling", is the theory of James and Lange [20]. To the other end, in Aristotelian theory, [12] emotion is a more or less intelligent way of conceiving a certain situation, determined by a desire (for example anger, the desire for revenge). Many of the modern theories involve both physiological and cognitive view toward the emotion. Some of them are concerned around the concept of *belief*, having in mind that if a person is in love, he or she must believe that the loved one has at least some virtues or attractions. It has also been noted that the analysis of emotion cannot be limited to the inner aspects of physiology and psychology, to visceral disturbances, sensations, desires and beliefs; emotion almost always has an outward aspect as well, its expression in *behavior*. From behaviorists point of view, emotion has expression in a behavior, so it can be defined as a distinctive pattern of behavior, sometimes recognized as *affective behavior*.

Among authors who have come to the idea to consider emotions as state evaluations is Brentano [11]. He follows the Aristotelian view of distinguishing emotions from cognitive and connative mental phenomena. In a sort of classification Brentano and Scheller [29] distinguish a class of emotions that have evaluative nature, from a class that has passionate nature. Sartre [28] and Solomon (e.g. in [12]) developed an evaluative theory in which emotions color the world with values.

What is the essence of viewing emotion as value judgement? It is a relation between emotions and evaluative beliefs. As a rule, what we feel about other people, events, and things in our lives generally indicates how we evaluate them. We value what we love, admire, envy, and feel proud of; we think ill of what we hate, fear, and find shameful or revolting. This way of thinking makes emotions logically dependent on evaluations. So emotions are (at least partly) evaluations. It can be said that evaluative theories compare pro- and con- emotional attitudes (liking, disliking, loving, hating, etc) and positive and negative value judgements.

Supported by these findings from psychology and philosophy, our theory of emotion considers emotion as a value judgement. According to this theory, an agent builds two types of emotions: 1) emotion induced by the internal state the agent is currently in, and 2) emotion about the previous behavior that lead to the situation the agent is currently in. The second type of emotion can be considered as *consequence driven belief*. It is a part of the logic of shame that anyone who feel ashamed must also hold a belief to the effect that he or she has acted wrongly. By analogy, having emotion of pride, one must have a belief that he or she has acted in a right way. Our theory of emotion has relation to the notion of valence, in appraisal theories of emotion [1],[22], [33].

3 Emotion Learning: Consequence Driven Systems Approach

Among three types of learning paradigms proposed within the theory of consequence driven systems [6] here we will consider the emotion learning paradigm, The theory considers three types of consequence learning agents, as shown in Figure 1.

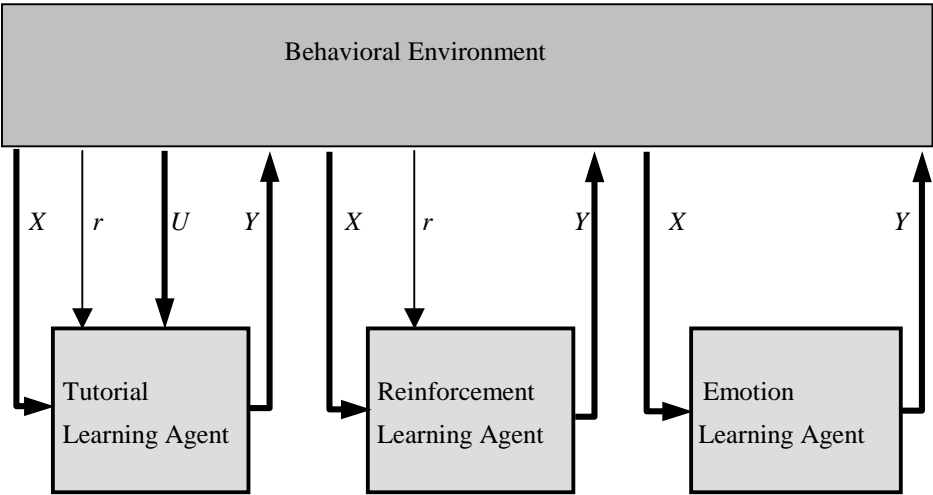


Figure 1. A taxonomy of consequence learning agents

As Figure 1 shows, the consequence driven systems theory deals with three types of learning systems: tutorial learning, reinforcement learning, and emotion learning. The tutorial learning agent from the environment receives a situation X , a (positive or negative) reinforcement r of what it did previously, but also an advice U , what to do in future. It produces behavior Y that affects the environment. The behavior can be a simple atomic action or a complex behavior containing sequences of actions. The other type of learning agent, the reinforcement learning agent did not receive any advice. It will adjust its behavior according only to the scalar value of the reinforcement. The third type of agent, the emotion learning agent will receive neither advice nor reinforcement; it will learn only from the received situation. Since there is no external reinforcement, it will build a scheme of a self-reinforcement based on its internal emotional value, $emotion(X)$, of the received situation X .

The theory argues [5],[6] that in order to build an emotion learning agent, the agent should posses 1) a genetic mechanism to receive initial, primary emotions from a genetic environment and 2) a emotion backpropagation mechanism (secondary reinforcement mechanism) which will develop emotions toward situations in the behavioral environment which in turn will develop the agent behavior by means of learning.

According to that, the emotion learning agent is a being which lives in two environments: behavioral and genetic. Contemporary research in Genetic Algorithms (e.g. [24]) and part of the research in Artificial Life (e.g. [21]) implicitly recognizes the genetic environment as standard feature of agents.

4 Emotion Learning Architecture

Here we will describe (Figure 2) our emotion learning architecture, named Crossbar Adaptive Array (CAA).. It consists of five main modules: crossbar associative memory module, behavioral environment interface module, emotion computation module, personality module, and genetic environment interface module.

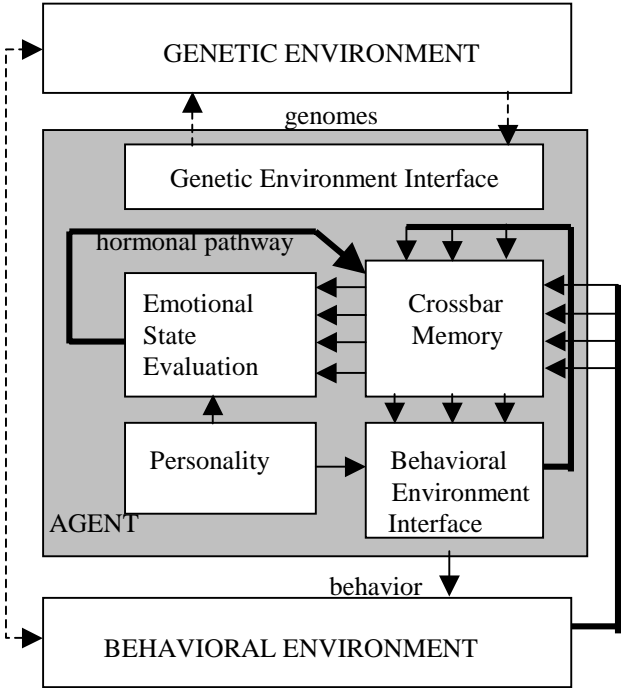


Figure 2. The CAA agent architecture

The principal component of the CAA architecture is its memory module, a connectionist weight matrix \mathbf{W} . Each component w_{aj} represents the *emotional value toward performing action a in the situation j* . It can be interpreted as tendency, disposition, expected reward, or by other terms depending on the considered problem to which CAA is used as solution architecture. From the emotional values for performing actions in situation k , CAA computes *emotional value v_k of being in the situation k* . Having v_k , CAA considers it as consequence of the action a performed previously in a situation j . So, it learns that performing a in j has as consequence emotional value v_k , and on that basis it learns the desirability of performing a in j . While the emotional values are computed column-wise, CAA, being in k , will compute the next action row-wise. In such a way, using crossbar computation over the crossbar elements, w_{aj} , CAA performs its *crossbar emotion learning procedure*, which has four steps:

- 1) state j : choose an action in situation j : $a_j = Afunc\{w_{sj}\}$
(let it be action a ; let the environment return situation k)
- 2) state k : feel the emotion being in state k : $v_k = Vfunc\{w_{sk}\}$
- 3) state j : learn the emotion toward a in j : $w_{aj} = Ufunc(v_k)$
- 4) change state: $j=k$; goto 1

where w_{sk} means that the computation considers all the possible desirability values of the k -th column vector. The functions $Afunc\{.\}$ and $Vfunc\{.\}$ are convenient to be defined as maximum selecting functions, but other functions can be used as well, provided they assure the convergence of the above procedure. For example, $Vfunc\{.\}$ can be in some cases a softmax function, i.e. sum of components, or some threshold based neural function. Both $Afunc\{.\}$ and $Vfunc\{.\}$ can be influenced by the personality parameters of the agent.

The learning rule, function $Ufunc(.)$, as used in CAA, is

$$w'_{aj} = w_{aj} + v_k \quad (1)$$

It is a simple learning rule, which just adds the computed emotional value v_k of the consequence situation, k , to the emotional value toward performing action a in situation j on which k is the consequence.

The learning procedure described above is actually an *emotion value backpropagation procedure* (secondary reinforcement learning procedure). Figure 3 emphasizes that using the Emotional Petri Nets concept [6]. As Figure 3 emphasizes, the states perform action selection and emotion backpropagation, but what is memorized, however, are the emotions associated to actions.

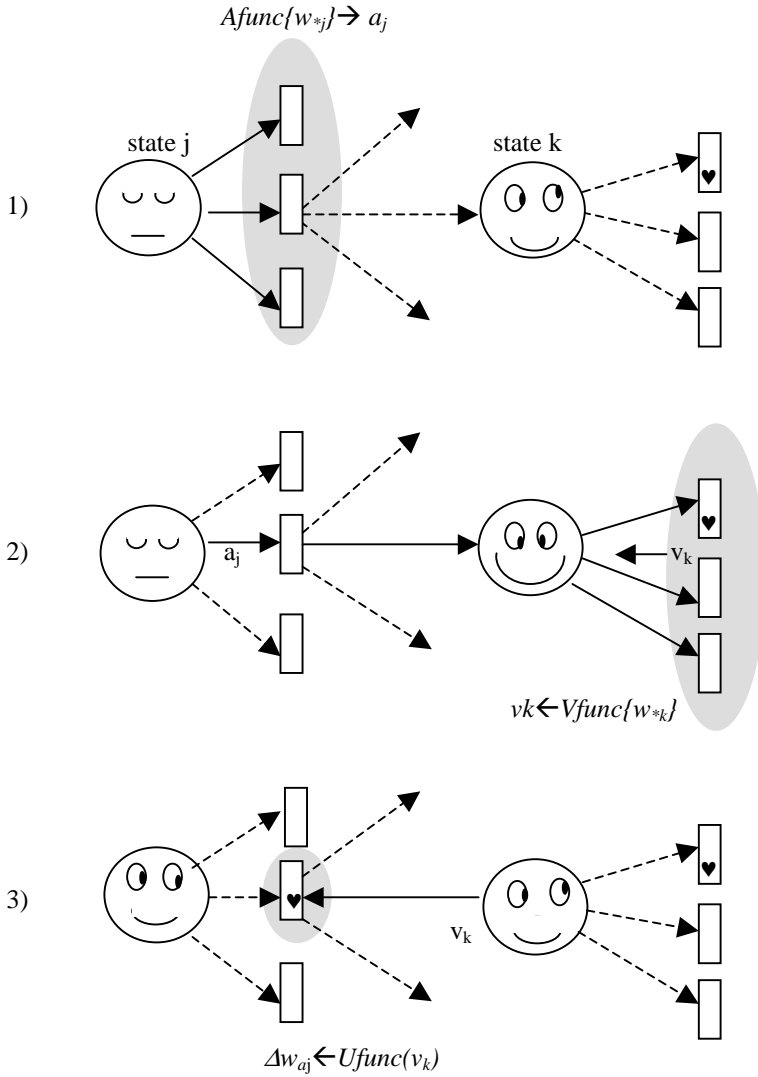


Figure 3. The CAA learning procedure represented by the Emotional Petri Nets

Consider the learning rule (1), and suppose that emotional value of v_j is zero, so no emotion is assigned to that situation. After the learning step is performed, the component w_{aj} of the situation j is incremented by the value v_k . So, after learning, the desirability of being in j is increased by v_k (in case v_k is negative it is decreased and potentially becomes an undesirable state). That means that the emotional value of the consequence state k is backpropagated to the causal state j . In such a way, after several runs, the CAA will learn a path toward the k -th state, and further to some goal state.

To describe the CAA philosophy, besides its crossbar computing matrix and its crossbar learning rule, we should describe how initial emotional preferences are defined. All the initial values are received from the genetic environment. The genetic environment is supposed to be a genotype reflection of the behavioral one. From the genetic environment CAA receives a string, denoted as genome string, or simply the CAA input genome. The input genome can have several chromosomes, but the most important is the memory chromosome. It contains initial values $\mathbf{W}(0)$ of the crossbar learning memory. So if the genetic environment properly reflects the desirable situations in the behavioral environment, then the genome string will properly inform the newborn CAA agent about the desirable (as well as the undesirable) situations in the behavioral environment. Having initial emotional values of the desirable situations, and using the emotional value backpropagation mechanism, the CAA will learn how to reach those situations. Having defined primary emotional values by the input genome, CAA learns a behavior in the environment by means of emotional value backpropagation using its learning rule

5 Developing an Emotion Learning Based, Trainable Goalkeeper Agent

Here we will give a short report on an application of this architecture in the domain of robot learning. As a robot learning task we will present training a RoboCup robot to take a role of a goalkeeper. Previously, the CAA architecture was used as controller in the task of emotional maze running [5] and the task of learning to balance a pole [8].

The RoboCup task is interesting, among other things, because the RoboCup contest is a new AI challenge, after the remarkable achievement motivated by the chess playing challenge. After building a computer program that is able to beat even the human world champion, the new challenge is stated as building a robot football (soccer) playing team that will beat the human world champion [2]. Various issues emerged around the new AI challenge, one of them being the issue of training, which the demonstration part of this work is situated in.

There are various roles in a robot football team, the basic ones being goalkeeper, defender and forward. The goalkeeper task is depicted in Figure 4. As Figure 4 shows, a simple strategy for a goalkeeper is to move along the goal line using some mechanism for line following, and try to minimize the angle α (see Figure 4). We recognized the similarity between this control problem and the inverted pendulum control problem. So we applied the CAA architecture as a controller of the goalkeeper, as we did for the inverted pendulum problem [6], [8] dividing the input space into 10 situations depending on the ball position and velocity.

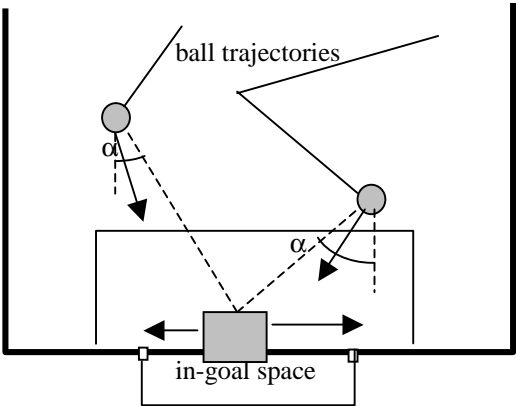


Figure 4. The goalkeeper task

As type of trainer we choose a random-shooting trainer that shoots the ball in random directions (angles) toward the goal side. The limitation is being put on the distance (minimal and maximal) from which the trainer shoots. The ball sometimes misses the goal (a miss-shoot), and sometimes bounces from the sidewalls (allowed by the RoboCup regulations). Some of the shots reach the goal, directly or after bouncing. Using CAA architecture, and defining the primary emotions such that receiving the goal produces a bad feeling, the robot is bound to learn to defend the goal against the ball.

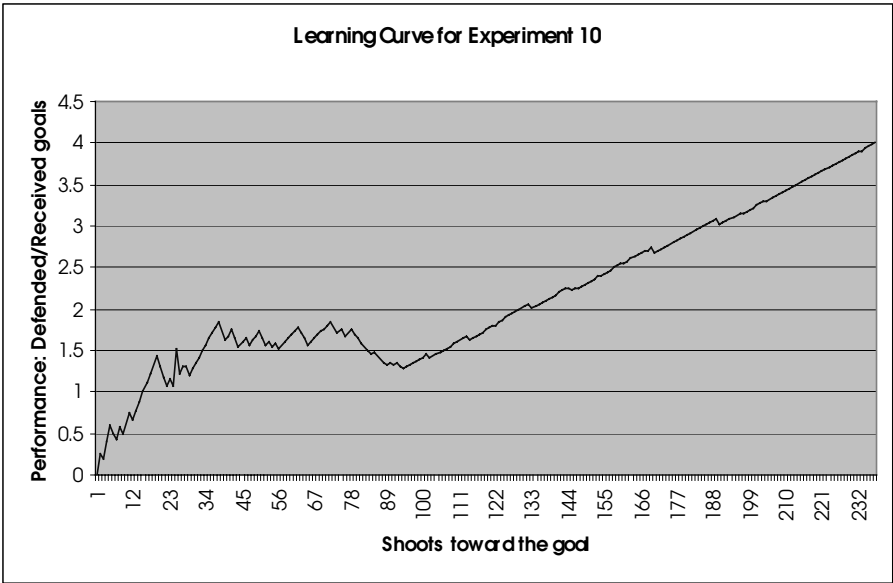


Figure 5. Learning curve for goalkeeper training

In a training process, one should distinguish a training phase and an exploitation phase (a real match). The threshold between the training and exploitation is defined by some performance measure that measures the skill and readiness of the goalkeeper to enter the exploitation phase. As performance coefficient we choose the ratio saved/received goals. We estimated that for a good human goalkeeper that ratio must be at least 4 (meaning 80% saved), so we trained the robot to achieve at least that performance. Figure 5 shows the training curve from a typical training experiment.

As Figure 5 shows, at the beginning the robot may save the goal by chance, but after several steps the training performance will drop and increase later due to the learned goalkeeping behavior. In this particular experiment, after about 90 shoots the robot has gained enough experience to *start behaving* as a goalkeeper. Its performance is getting better gradually. However, although it is trained to recognize most of the situations after 90 training trials, some of the situations are learned later during the training. The training in this experiment is carried out until the training performance ratio reaches the value 4. It is assumed that this goalkeeper is now well trained and can enter its exploitation phase. Note that the learning curve in Figure 5 shows its saturation as a linearly increasing tendency. A standard saturation toward the value of 1 is seen using the performance ratio saved/(received+saved) goals.

We carried out a number of successful experiments on our Java simulator. The success was measured in a sense that all experiments achieved a performance ratio greater than 4. Actually, in our training experiments we achieved a ratio between 4 and 10 in less than 500 training shots (including misses).

6 Related Work

Being a subject of interest in psychology for centuries, the issue of emotion in cognitive science and AI seems to be invoked by Simon [31]. Some early works on emotion and feelings in agent architectures are the conceptual architecture of mind connecting emotions to motives proposed in [32], and the connectionist learning architecture proposed in [5] which used affective representation of emotional states using facial expressions. Several cognitive science related books cover the subject of emotions and emotion based agents, including [23], [16], [3], [6], [18], [22], [26]. A review on contemporary emotion research effort is given in [15].

Besides the relation to emotion-based agents and architectures, this work is related to reinforcement learning, especially to Q-learning [37], [4], [8]. The work [5] introduced three crucial notions to reinforcement learning: 1) the notion of state evaluation, 2) the term “delayed reinforcement learning” instead of the previously used term “assignment of credit”, and 3) the learning procedure in which an agent should compute state values but learn and remember the action values (Figure 3). The author believes that this is actually what is currently meant in AI under the notion of Q-learning [27]. It is out of scope of this paper to elaborate this further, but one point should be emphasized, which is relevant to emotion research and affective computing approach: the notion of emotion and feeling was present from the very beginning of introduction of state evaluation concept in reinforcement learning.

Also, learning to play football is gradually gaining attention within the RoboCup community, and is heavily related to this work. The work within the GMD RoboCup

team [9] and the dual dynamics approach [10] provided preconditions for this work and the described application. There have been several reports on training a football playing robot. An interesting approach has been reported in [34]. A robot was trained for a shooting role, where a supervised pattern classification method was implemented. A human expert selected a training set of various possible shooting situations, and assigned a desired shoot for each situation. It was shown that a successful training is achieved using this method. Another approach has been reported for passing-shooting pair from the Osaka team. It was mentioned [2] that the Osaka team used reinforcement learning for training the goalkeeper too, but we have not yet had the opportunity to study the description of this training.

7 Conclusion

This paper considers the problem of how to define the concept of emotion in robots and how to implement that concept in the learning process. The notions of artificial emotion and emotion learning are associated to this effort. The discussion is situated within the theory of consequence driven systems, where the emotion learning agent is assumed to be an agent that has ability to learn without an external reinforcement; it had been argued that such an agent can learn only if it develops an internal evaluation mechanism, and indeed, emotions are such a mechanism. Basing on that observation, and finding support in evaluative theories of emotion proposed in psychology and philosophy, this research develops an emotion learning architecture, based on a connectionist crossbar adaptive array. That architecture uses a learning rule which indeed implements emotion in its structure.

The paper is believed to contribute toward computational theories of emotion. It presents a working theory and model which could influence building emotion-based agents. It also presents a psychological and philosophical grounds for the approach taken. In its application illustration part, this work briefly reports on an attempt to train a goalkeeper robot, which is believed to be an interesting application of the emotion-learning concept. The clear message of the paper is that the approach using emotion as value judgement could be a fruitful approach of looking for emotions in artificial agents.

Acknowledgement. The author wishes to express gratitude to Prof. Dr Thomas Christaller and Dr Herbert Jaeger for their invitation to work within the GMD AiS BE Group, RoboCup Team, and for excellent working conditions and support that lead to the experimental results reported in this work. Also the author wishes to thank Dr Bernd Mueller and Peter Schoell for advices and support.

References

- 1 Arnold M. 1960. *Emotion and Personality*. Columbia University Press.
- 2 Asada M., Kitano H., Noda I., Veloso M. 1999. RoboCup: Today and tomorrow - What we have learned. *Artificial Intelligence* 110: 193-214
- 3 Balkenius C. 1995. *Natural Intelligence in Artificial Creatures*. Lund University, 1995
- 4 Barto A. 1997. Reinforcement learning. In O. Omidvar and D. Elliot (Eds.) *Neural Systems for Control*, p. 7-29, Academic Press
- 5 Bozinovski S. 1982. A self-learning system using secondary reinforcement. In R. Trapp (Ed.) *Cybernetics and Systems Research* pp.397-402, North-Holland
- 6 Bozinovski S. 1995. *Consequence Driven Systems*. Gocmar Press
- 7 Bozinovski, S., Stojanov G., Bozinovska, L. 1996. Emotion, embodiment, and consequence driven systems. *Proc AAAI Fall Symposium on Embodied Cognition and Action*, p. 12-17, AAAI Press
- 8 Bozinovski S. 1999. Crossbar Adaptive array: The first connectionist network that solved the delayed reinforcement learning problem. In A. Dobnikar, N. Steele, D. Pearson, R. Albrecht (Eds.) *Artificial Neural nets and Genetic Algorithms* pp. 320-325, Springer Verlag
- 9 Bozinovski S., Jaeger H., Schoel P. 1999. Engineering goalkeeper behavior using an emotion learning method. In S. Sablatnoeg, S. Enderle (Eds.) *Proc RoboCup Workshop, KI99: Deutsche Jahrestagung fuer Kuenstliche Intelligenz*, pp. 48-56, Bonn
- 10 Bredenfield A., Christaller T., Goehring W., Guenter H., Jaeger H., Kobialka H-U., Ploeger P-G., Schoell P., Sieberg A., Verbeek C., Wilberg J. 1999. Behavior engineering with "dual dynamics" model and design tools. *Proc ICAI-99 RoboCup Workshop*, Stockholm
- 11 Brentano F. 1889. *On the Origin of our Knowledge of Right and Wrong*. Humanities Press, 1969
- 12 Calhoun C., Solomon R. 1984. *What is an Emotion*, Oxford University Press
- 13 Canamero D., Numaoka C., Petta P. (Eds.) 1998a. *Grounding Emotions in Adaptive Systems*. Workshop at the Simulation of Adaptive behavior Conference, Zurich.
- 14 Canamero D. (Ed.) 1998b. *Emotional and Intelligent: The Tangled Knot of Cognition*. AAAI Fall Symposium, AAAI Press
- 15 Castelfranchi C. 1999. Affective appraisal vs cognitive evaluation in social emotions and interactions (this volume)
- 16 Damasio A. 1994. *Descartes' Error, Emotion, reason and the Human Brain*, Grosset/Putnam
- 17 Gadanho S., Hallam J. 1998. Exploring the role of emotions in autonomous robot learning. *Proc AAAI Fall Symposium on Emotional and Intelligent: The Tangled Knot of Cognition*, p.84-89, AAAI Press
- 18 Goleman D. 1966. *Emotional Intelligence: Why it can matter more than IQ*. Bloomsbury
- 19 Jaeger H., Christaller T. 1998. Dual dynamics Designing behavior systems for autonomous robots. *Artificial Life and Robotics* 2: 108-112
- 20 James W., and Lange C. 1885. *The Emotions*. Williams and Wilkins, 1922
- 21 Langton C. 1989. Artificial life. In C. Langton (ed.) *Artificial Life*, Volume VI of SFI Studies in the Science of Complexity, pp.1-47, Addison Wesley
- 22 LeDoux J. 1996. *The Emotional brain*. Simon and Schuster
- 23 Minsky M. 1988. *The Society of Mind*. Simon and Schuster
- 24 Mitchell M. 1995. Genetic algorithms: An overview. *Complexity* 31-39
- 25 Paiva A., Martinho C. 1999. (Ed.) *Proc Workshop on Affect in Interactions*, Siena
- 26 Picard R. 1997. *Affective Computing*. The MIT Press
- 27 Russel S., Norvig P. 1995. *Artificial Intelligence*. Prentice Hall

- 28 Sartre J-P. 1939 . A Sketch of a Phenomenological Theory. from *Emotions*. Philosophical Library. 1948
- 29 Scheler M. 1916, *Formalism in Ethics and Non-Formal Ethics of Values*. Northwestern University Press, 1973
- 30 Seif El-Nasr M., Ioerger T., Yen J. 1998. Learning and emotional intelligence in agents. *Proc AAAI Fall Symposium on Emotional and Intelligent: The Tangled Knot of Cognition*, p. 150-155, AAAI Press
- 31 Simon H. 1979. Motivational and emotional controls of cognition. In *Models of Thought*, 29-38, Yale University Press
- 32 Sloman A., Croucher M. 1981. Why robots will have emotions. *Proc 7th Int ICAI*, 197-202
- 33 Staller A., Petta P. 1998. Towards a tractable appraisal based architecture for situation cognizers. *Proc Workshop on Grounding Emotions in Adaptive Systems*, Zurich. ,
- 34 Tambe M., Adibi J., Al-Onaizan Y., Erdem A., Kaminka G., Marsella S., Muslea I. 199. Building agent teams using an explicit teamwork model and learning. *Artificial Intelligence* 110: 215-239
- 35 Velasquez J. 1998. Modeling emotion-based decision making. *Proc AAAI Fall Symposium on Emotional and Intelligent: The Tangled Knot of Cognition*, p. 164-169, AAAI Press
- 36 Velasquez J. (Ed.) 1999 *Proc Workshop on Emotion based agent architectures*, Seattle, 1999
- 37 Watkins C. 1989. Learning from delayed rewards. PhD Thesis, King's College, Cambridge

Integrating Models of Personality and Emotions into Lifelike Characters

Elisabeth André, Martin Klesen, Patrick Gebhard, Steve Allen, and Thomas Rist

DFKI GmbH,
Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany,
{andre,klesen,gebhard,allen,rist}@dfki.de

Abstract. A growing number of research projects in academia and industry have recently started to develop lifelike agents as a new metaphor for highly personalised human-machine communication. A strong argument in favour of using such characters in the interface is the fact that they make human-computer interaction more enjoyable and allow for communication styles common in human-human dialogue. In this paper we discuss three ongoing projects that use personality and emotions to address different aspects of the affective agent-user interface: (a) *Puppet* uses affect to teach children how the different emotional states can change or modify a character's behaviour, and how physical and verbal actions in social interactions can induce emotions in others; (b) the *Inhabited Market Place* uses affect to tailor the roles of actors in a virtual market place; and (c) *Presence* uses affect to enhance the believability of a virtual character, and produce a more natural conversational manner.

1 Introduction

A growing number of research projects in academia and industry have recently started to develop lifelike agents as a new metaphor for highly personalised human-machine communication. A strong argument in favour of using such characters in the interface is the fact that they make human-computer interaction more enjoyable and allow for communication styles common in human-human dialogue. Our earlier work in this area concentrated on the development of animated presenters that show, explain, and verbally comment textual and graphical output in a window-based interface. Even though first empirical studies have been very encouraging and revealed a strong affective impact of our Personas [28], they also suggest that simply embodying an interface agent is insufficient. To come across as believable, an agent needs to incorporate a deeper model of personality and emotions, and in particular, directly connect these two concepts.

The German Research Centre for Artificial Intelligence (DFKI) recently started three new projects to advance our understanding of the fundamental technology

required to drive the social behaviour of agents. This initiative has been timed to catch the current wave of research and commercial interest in the field of lifelike characters [1] and affective user interfaces [29]. The i3-ese project Puppet promotes the idea of a virtual puppet theatre as an interactive learning environment to support the development of a child's emotional intelligence skills. The second project features an Inhabited Market Place in which personality traits are used to modify the characters' roles of virtual actors in sales presentations. The Presence project uses an internal model of the agent's (and possibly the user's) affective state to guide the conversational dialogue between agent and user. Although all three projects rely on a more or less similar approach towards modelling emotions and personality traits, there are variations with regard to the underlying user-agent(s) relationship(s), the factors that influence an agent's emotional state, the complexity of the underlying model of emotions and the way in which emotions and personality traits are made observable. The following sections provide short overviews of the three projects and discuss their affective nature in more detail.

2 Basic Concepts

One of the first challenges we must face when attempting to use *affect* within our architectures, is to recognise the fact that the term does not refer to a well-defined class of phenomena clearly distinguishable from other mental and behavioural events. Affect is used within the literature to describe the class of motivational control states which result from valenced reactions to objects and events - these include emotions, mood, and arousal. Therefore the only generalisation we can really make about affect is that it must contain at least the two attributes of activation and valence. The different classes of *affective states* can further be differentiated by *duration*, *focus*, *intensity*, and *expression/effect* - emotions tend to be closely associated with a specific event or object and have a short duration, whereas mood is more diffuse and of longer duration. Within the context of this paper, we define *personality* as "the complex of characteristics that distinguishes an individual or a nation or group; especially the totality of an individual's behavioural and emotional characteristics", and *emotion* as "affect that interrupts and redirects attention (usually with accompanying arousal)" [26].

Although there is no consensus in the nature or meaning of *affect*, existing theories and models of personality and emotion can still play an useful role in enhancing user-agent interaction - even though they do not capture the *affective* phenomena in its entirety. As a starting point for our work, we have taken the Five Factor Model (FFM) [18] of personality, and the Cognitive Structure of Emotions model (OCC - Ortony, Clore and Collins) [20]. These models are readily amenable to the intentional stance, and so ideally suited to the task of creating concrete representations/models of personality and emotions with which to enhance the illusion of believability in computer characters.

Emotions: The OCC model of emotions provides a classification scheme for common emotion labels based on a valence reaction to events and objects in the light of agent goals, standards, and attitudes. The OCC model is a model of causation, and

will be used within both Presence and Puppet to determine the affective state of the character in response to events in the environment (see also [6] and [23]).

Personality: The FFM is a purely descriptive model, with the five dimensions (*Extraversion*, *Agreeableness*, *Conscientiousness*, *Neuroticism*, and *Openness*) being derived from a factor analysis of a large number of self- and peer reports on personality-relevant adjectives. The descriptive nature of the FFM gives us an explicit model of the character's personality, and in turn, allows us to concentrate on using the affective interface to directly express those traits (which offers the interesting possibility of attempting to recreate the character's personality traits from an analysis of the emergent social interaction). Furthermore, as we are focusing on social interactions, we can concentrate on the traits of *extraversion* (Sociable vs. misanthropic; Outgoing vs. introverted; Confidence vs. timidity) and *agreeableness* (Friendliness vs. indifference to others; A docile vs. hostile nature; Compliance vs. hostile non-compliance) - although we will also use *neuroticism* (Adjustment vs. anxiety; Level of emotional stability; Dependence vs. independence) to control the influence of emotions within our characters.

In addition to generating *affective states*, we must also express them in a manner easily interpretable to the user (which in the case of the Puppet project will be young children). Personality and emotions can be conveyed in various ways. According to empirical studies, extravert characters use more direct and powerful phrases than introvert characters [8], speak louder and faster [25] and use more expansive gestures [9]. Furthermore, the rendering of dialogue acts depends on an agent's emotional state. Effective means of conveying a character's emotions include acoustic realisation, body gestures and facial expressions [5]. While these studies seem directly applicable to anthropomorphic agents like the Presence Persona, it is not clear to what extent they apply to animals with anthropomorphic features such as the characters in the Virtual Puppet theatre.

In all three projects, personality and emotions are used as filters to constrain the decision process when selecting and instantiating the agent's behaviour. For instance, we might define specific behaviours for extravert characters in a certain emotional state. However, there are other (affective) states we would like to convey that are not simply the result of an affective appraisal (as in the OCC model), or easily derived from personality traits - i.e. fatigue, boredom, and hunger. To model these states, we will mimic our character's active body state with motivational drive mechanisms to provide the affective input signals.

3 The Role of Affect in Puppet

The objective of the Puppet project is to develop and investigate the value of a new virtual reality environment, the Virtual Puppet Theatre, (VPT), based on a theoretical framework of "learning through externalisation" [24]. Deploying user-controlled avatars and synthetic characters in the child's own play production, the children have to distinguish and master multiple roles in their interaction with the system, e.g. that of a director, an actor and an audience with the main activities producing, enacting and reflecting respectively. Within this process the children should gain a basic

understanding on how different emotional states can change or modify a character's behaviour and how physical and verbal actions in social interaction can induce emotions in others. These emotional intelligence skills are important for us with respect to the early learning goals: "social role decentring" and theory of mind. Our approach is similar to [10] which allows children to direct a puppet's mood, actions and utterances in interactive story-making and to [15] where children may induce some changes in their characters emotional state besides selecting a character's actions.

3.1 Application Domain

For our first prototype (VPT1) developed for children at the age of 5-6, we decided to model a farmyard as a co-habited virtual world, in which the child's avatar (e.g. the farmer) and a set of synthetic characters (pigs, cows, etc.) can interact with each other. Fig. 1 shows a screenshot of the virtual 3D environment which was developed by our project partners from the Laboratory of Image Analysis at the University of Aalborg. Our characters are designed to exhibit both physical and verbal behaviour. We do not try to model "real" animals but make them more cartoon-like instead.



Fig. 1. 3D Prototype of the farmyard scenario

For the communication between the avatar and a character we will use a simple speech-act based dialogue model and a set of pre-recorded utterances. The agents are equipped with virtual sensors and effectors which connect them to the 3D virtual environment and controlled by an agent architecture that integrates deliberative (goal-driven) and reactive (data-driven) planning. To foster the above mentioned emotional skills we provide two distinct sets of interfaces which can be used by the child to

control a character's behaviour. A *body control interface* which gives full control over the movement of the selected character and a *mind control interface* which allows to change the character's emotional state thus biasing the behaviour in some direction without specifying the actual motion pattern. The mind control interface is icon-based with prototypical facial expressions for the modelled emotion types. The design of the interface is based on empirical studies [16]. Selecting an utterance is a shared effort between the child and an adult helper. Utterances with the same speech-act assigned to them are treated as equivalent by the system. Similar to the systems described in [10] and [17], we separate the high-level behaviour planning and affective reasoning (the "mind") from the animation planning and control modules (the "body"). The first is done by the agent architecture as described in the next section and the latter lies within the responsibility of the 3D virtual environment.

3.2 Emotions and Personality

The agent architecture used for the high-level behaviour planning and affective reasoning consists of a knowledge base, a plan library, an interpreter and an intention structure. The knowledge base is a database that contains the world model (the "beliefs") of a character. The plan library is a collection of plans that an agent can use to achieve its goals and the intention structure is an internal model of the current goals (the "desires") and instantiated plans (the "intentions") of that agent. Within this architecture we can have multiple active goals and multiple plans for each goal. Conflict resolution, plan selection, instantiation and execution are handled by the interpreter (see Fig. 2).

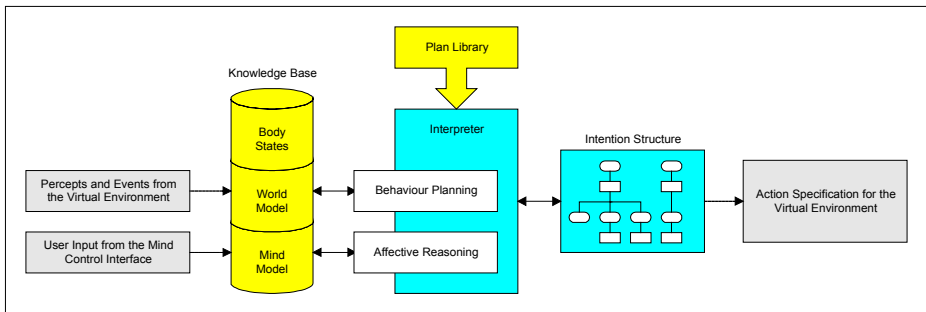


Fig. 2. Agent architecture for VPT1

There are two types of input which can influence an agent's behaviour planning and affective reasoning: percepts and events from the virtual environment; and user input from the mind control interface. The first is used to update the agent's world model (which also influences the affective reasoning process) and the second to directly change its affective state encoded in the mind model. The interpreter will modify and replace all ongoing behaviours affected by these changes and might even instantiate new ones if they are better suited, i.e. have a higher utility, to express the puppet's mental state in the current context. To increase the life-likeness of a character we also introduce body states (fatigue, boredom, hunger) which are

represented within the knowledge base and regularly updated by the interpreter. The body states act as motivational drives that impel the agent into action by activating the appropriate behaviour (sleeping, playing, eating and drinking), which is then carried out in a character-specific way (if a dog is hungry it will go to the farmhouse whereas a cow will start grazing). The output of the planning processes is an action specification for the virtual environment. It contains appropriate mark-ups (e.g. for the facial expression) taking into account the current emotional state.

As a starting point for the mind model capturing the personality and affective states of our characters we use the OCC model. In Puppet it is particularly important that we can express these states in a manner easily interpretable by young children. We therefore decided to model the emotion types *Anger*, *Fear*, *Happiness* and *Sadness* based on evidence suggesting their universality [5] and the fact that there are distinctive facial expressions which can be interpreted properly by children of age 4-8 [22]. Emotions are primarily conveyed by facial expressions and the selection of appropriate sounds (a cat will purr if it is happy and hiss if it is angry). They are either computed by emotion generating rules according to the OCC model or directly manipulated by the child through the mind control interface (see Fig. 2).

To model personality, we adopt the FFM, but reduce it to the dimensions extroversion and agreeableness because they determine to a large extent how an agent will behave in social interactions [12]. In addition we specify for each character a set of preferences (e.g. the dog likes bones) and long term goals. Most characteristics are tailored for each character to give them unique pseudo personalities. This means that we can only partially rely on the same high-level behaviour to convey personality features (e.g. greet another character and start playing if you are extrovert and agreeable) and that we have to devise character-specific ones otherwise.

Puppet offers a variety of different user-agent(s) relationship(s). In “enacting mode” the child uses an avatar to interact with other characters in the scene. This is similar but not identical to Presence where the user interacts with the Persona through a set of input devices. The second mode, the child playing the role of an audience by observing the interaction of two or more autonomous agents has its equivalent in the Inhabited Market Place where the user observes a dialogue performed by a team of characters. However there is a third distinct user-agent relationship in Puppet, namely that of the child being a director, i.e. controlling the behaviour of *all* characters in the scene. This is similar to make-believe play with physical puppets during childhood in which the child takes on a series of roles. The difference is that the animals in our scenario are semi-autonomous, i.e. they take directions (e.g. the child can force the puppets to do or say something or change its affective states) that bias but not completely specify their behaviour. How an agent will (re)act in a specific situation also depends on its internal body states and personality features. This mode could provide valuable insights because we can observe when and how the children change the emotional state of a character, something that is not so easy to infer in conventional make-believe play.

Over the next few months, our project partners at COGS will evaluate a group of 5-6 year old children before and after they played with the VPT. We hope that their findings will validate our assumption that the children’s emotional intelligence skills will be improved by constructing simple models of the virtual puppets minds. It will be also interesting to see how the ability to take the subjective view of different

characters (as an actor) and to direct their behaviour (in the role of a director) will increase their understanding of the dynamics in social interactions, especially how emotions influence these interactions.

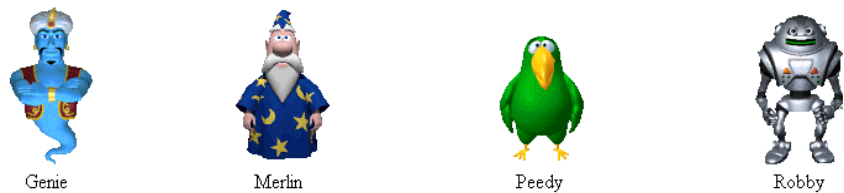
4 The Role of Affect in The Inhabited Market Place

The objective of the Inhabited Market Place is to investigate sketches, given by a team of lifelike characters, as a new form of sales presentation. The basic idea is to communicate information by means of simulated dialogues that are observed by an audience. The purpose of this project is not to implement a more or less complete model of personality for characters, such as a seller and a customer. Rather, the demonstration system has been designed as a testbed for experimenting with various personalities and roles.

4.1 Application Domain

As suggested by the name, the inhabited market place is a virtual place in which seller agents provide product information to potential buyer agents. For the graphical realisation of the emerging sales dialogues, we use the Microsoft Agent package [19] that includes a programmable interface to four predefined characters: Genie, Robby, Peedy and Merlin. To enable experiments with different character settings, the user has the possibility of choosing three out of the four characters and assigning roles to them (see Fig. 3). For instance, he or she may have Merlin appear in the role of a seller or buyer. Furthermore, he or she may assign to each character certain preferences and interests.

Select the agents and their personality:



SELLER Genie ▾		BUYER1 Peedy ▾		BUYER2 Merlin ▾	
Agreeableness	Extraversion	Agreeableness	Extraversion	Agreeableness	Extraversion
<input checked="" type="radio"/> agreeable	<input type="radio"/> extravert	<input type="radio"/> agreeable	<input type="radio"/> extravert	<input checked="" type="radio"/> agreeable	<input checked="" type="radio"/> extravert
<input type="radio"/> neutral	<input type="radio"/> neutral	<input type="radio"/> neutral	<input type="radio"/> neutral	<input type="radio"/> neutral	<input type="radio"/> neutral
<input type="radio"/> disagreeable	<input checked="" type="radio"/> introvert	<input checked="" type="radio"/> disagreeable	<input type="radio"/> introvert	<input type="radio"/> disagreeable	<input type="radio"/> introvert

Fig. 3. Dialog for character settings

The system has two operating modes. In the first mode, the system (or a human author) chooses the appropriate character settings for an audience. The second mode allows the audience to test various character settings itself. Fig. 4 shows a dialogue between Merlin as a car seller and Genie and Robby as buyers. Genie has uttered some concerns about the high running costs which Merlin tries to play down. From the point of view of the system, the presentation goal is to provide the observer – who is assumed to be the real customer – with facts about a certain car. However, the presentation is not just a mere enumeration of the plain facts about the car. Rather, the facts are presented along with an evaluation under consideration of the observer's interest profile.

4.2 Emotions and Personality

In the sales scenario, the role of the system may be compared with that of a screen writer who produces a script for the actors of a play. The script represents the dialogue acts to be executed by the individual agents as well as their temporal order. To automatically generate such scripts, we use a plan-based approach similar to that in Puppet. Knowledge concerning the generation of scripts is represented by means of plan operators.

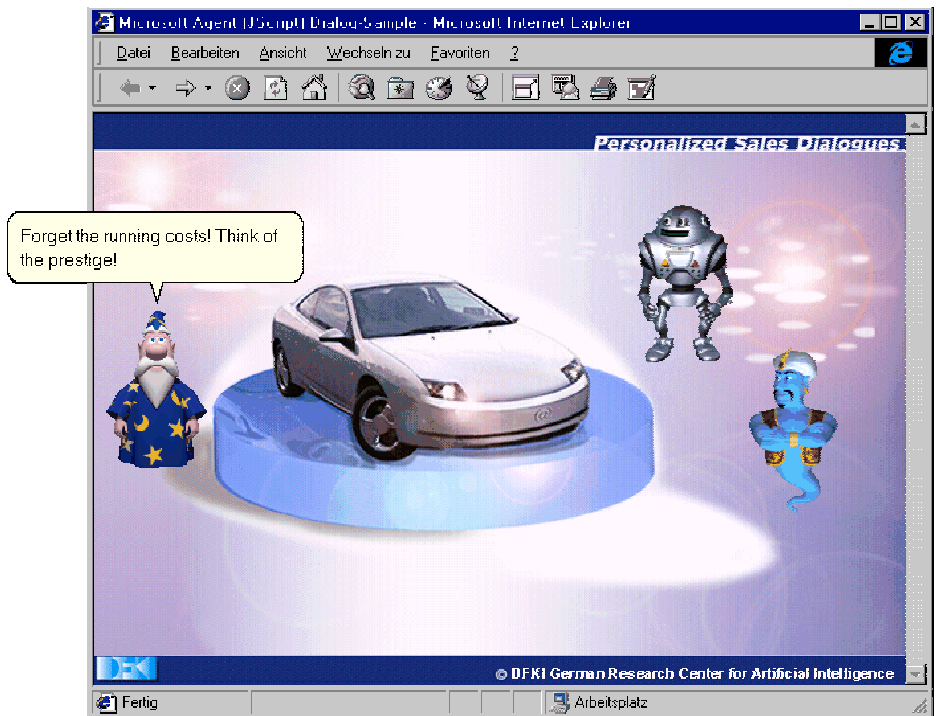


Fig. 4. Car sales dialogue example

As in Puppet, we decided to model two personality factors, namely: *extraversion* and *agreeableness*. In the first version, we concentrate just on one dimension of emotive response, namely valence [13], with the possible values: positive, neutral, and negative.

In the sales scenario, emotions are essentially driven by the occurrence of events. The events in the sales scenario are the speech acts of the dialogue participants that are evaluated by the characters in terms of their role, personality traits, and individual goals. The goals in particular determine the desirability of events, for example, a buyer will be displeased if he is told that a relevant attribute of a car (e.g. power windows) is missing for a dimension that is important to him (e.g. comfort). In this scenario, we do not deal with emotion structures and emotion generating rules explicitly (e.g., see [6]) but rather connect the scenario-specific dialogue acts (e.g., DiscussValue, PositiveResponse, InformIf) to the relevant animation sequences and utterance templates by using the current internal state of the character as an additional constraint in the behaviour selection mechanism. This approach is similar to that of Lester and colleagues [14], where pedagogical speech acts drive the selection and sequencing of emotive behaviours.

Personality in the Inhabited Market Place is essentially conveyed by the choice of dialogue acts and the semantic and syntactic structure of an utterance. Emotions in this scenario are expressed by facial expressions and the syntactic structure of an utterance. Since the Microsoft Agent Programming tool does not allow for detailed intonational mark-ups, we do not convey emotions by acoustic realisation in this scenario.

First informal system tests were encouraging. Even though it was not our intention to make use of humour as e.g. the authors of the Agneta & Frida system [11], people found the generated dialogues entertaining and amusing. Furthermore, people were very eager to cast the agents in different roles in order to find out the effect this would have on the generated presentations. These observations suggest that people would possibly learn more about a subject matter because they are willing to spend more time with a system.

5 The Role of Affect in Presence

The Presence project will use lifelike characters as virtual receptionists/infotainers/accompanying guides for visitors to the German Research Centre for Artificial Intelligence (DFKI). Here we will explore the hypothesis that using an explicit affective model (of both agent and user) to guide the presentation strategies used in the human-agent conversational dialogue will (a) create a more natural and intuitive user interface (by tailoring the conversation to an individual person); (b) provide the user with an engaging and enjoyable experience; and (c) enhance the believability of virtual characters.

The Presence project addresses a number of specific problem areas: (a) flexible integration of multiple input (speech, mouse, keyboard and touch-screen) and output (text, pictures, videos and speech) devices. The architecture must be intelligent enough to adapt to the different affective modes of communication in the different

application domains - i.e. no speech inflection in the remote domain; (b) the development of a high-level descriptive language for character definition, based on personality traits to allow easy customisation of the agent; (c) the combination of computational models of personality and emotion with planning techniques to guide the interaction of a lifelike character presenting material to visitors both locally and/or remotely over the world wide web; and (d) explore the possibility of tailoring the agent-user interaction to an individual user by inferring the user's affective state (see also [3]).

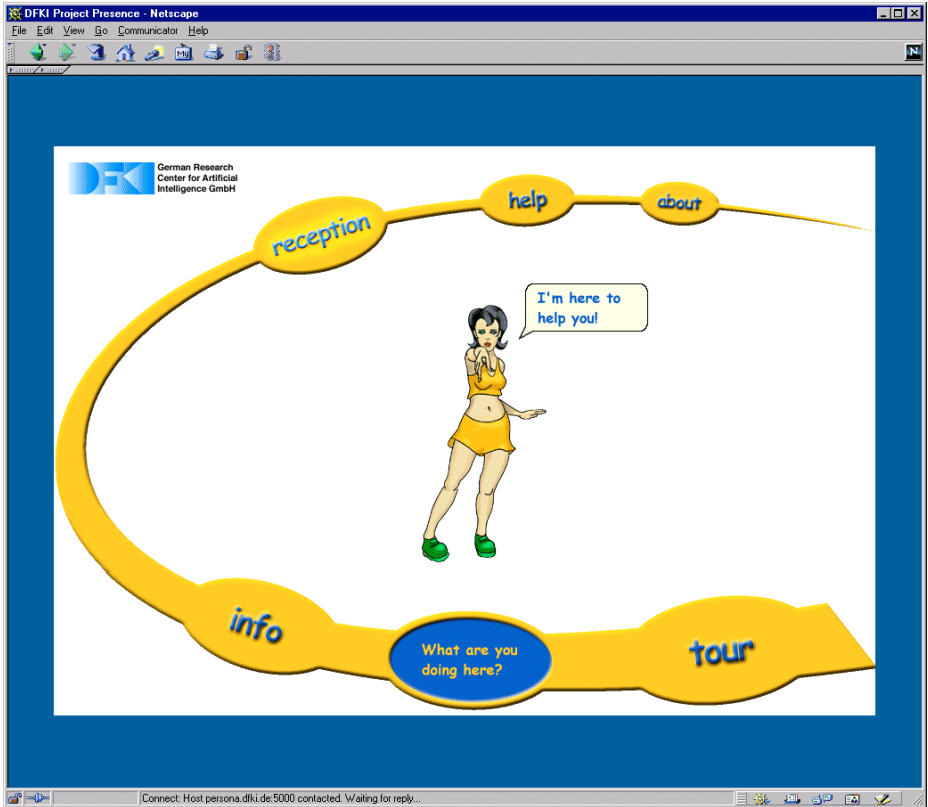


Fig. 5. Screenshot of the Presence Prototype

5.1 Application Domain

One of the major design considerations of the project is the requirement for a flexible architecture capable of coping with multiple application domains. Domains are defined by the scope of information they cover and the general interaction behaviour of the lifelike character. Domains are furthermore decomposed into hierarchically structured dialogue topics (i.e. Welcome/Small-talk, Helpdesk, Info/Tour) which guide the conversational thread. Our domains are:

- *Receptionist (Entrance Hall)*: Here the Presence system will run on an infoterminal within the DFKI entrance hall, and will welcome visitors, business partners, and student to the institute. The *virtual receptionist* will answer questions on a wide range of dialogue topics covering news, research projects, and people within the DFKI. The receptionist will also be capable of initiating conversations and informally introducing the various dialogue topics. The actual conversational thread will be guided by both the user responses, and the modelled affective state of the agent.
- *Infotainment (Remote)*: The remote infotainment domain will utilise the same underlying conversational modes as the local receptionist domain. However, the constraints imposed by the restricted bandwidth of the internet will place an additional requirement on the system to make best use of the available input and output resources. Our system must therefore be intelligent enough to cope with a varying communication channel with the user, i.e. one in which the affective channel of speech inflection may no longer be available.
- *Guide (Portable)*: Within the portable application domain, the system's primary role will be to guide the user through the building, i.e. the user can ask the system how to reach a lab or office. However, instead of remaining a passive guide, the Presence system will take advantage of the infra-red channel of a palm computer to provide a low bandwidth link to the server - thus allowing the system to update the user with DFKI internal news, or to signal the beginning of lectures, meetings or talks. The portable domain will provide a real challenge to convey affective information in such an impoverished environment.

5.2 Emotions and Personality

The Presence system model extends the PPP animated presentation agent architecture developed at DFKI [2] - most notably with enhanced input and output modalities for *affective* communication, and an *Affective Reasoning Engine* for *affective state* recognition and modelling. In line with recent research in affective computing [21], we use two *affective* information processing channels (see Fig. 6).

Our two affective channels are used to model *primary* and *secondary* emotional states - a cognitive classification scheme based on the information-level processes involved in the generation of the emotion types (see [4] and [27]). *Primary* emotions (i.e. being startled, frozen with terror, or sexually stimulated) are generated by innate neural machinery in the reactive cognitive layer - centred around the human limbic system. *Secondary* emotions can either arise: (a) through learned associations between categories of objects and situations attended to by deliberative thought processes on the one hand, and primary emotions, on the other [4]; or (b) in response to the deliberative planning process itself (when relevant risks are noticed, progress assessed, and success detected) [27]. Secondary emotions therefore require a deliberative cognitive layer in addition to the reactive machinery of primary emotions.

As emotion types (happiness, sadness, envy, etc.) often exhibit very different characteristics (i.e. varying degrees of cognitive richness and/or expressiveness) depending on the layers of the cognitive architecture involved in the emotion process, we felt that it was important to distinguish between these two classes of emotional

state in Presence. For example, fear can be generated: (a) as an innate response to a situation/event in the external environment – a *primary* emotion; (b) by cognitively identifying a potential future threat – *secondary* emotion; or (c) as a perturbant state when we repeatedly attempt to reassure ourselves that the threat is not real – a *tertiary* emotion [27].

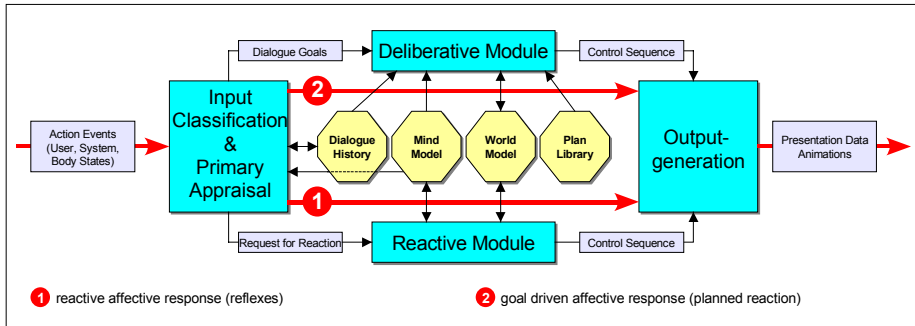


Fig. 6. Conceptual view of the Presence system showing two affective pathways

Each class of fear has its own physiological characteristics (with primary emotional states eliciting the strongest physiological response), and hedonistic tone (with tertiary emotional states being the most cognitive in nature). In Presence, *primary* emotions are modelled using simple reactive heuristics, whereas the more complex *secondary* emotions are modelled by the deliberative *Affective Reasoning Engine* according to the OCC model (making the finer grade distinction of *tertiary* emotions - secondary emotions that temporarily reduce attentive self-control - was not considered necessary in our application domain).

Emotions and personality traits weave an intricate web within the Presence system. Emotions are primarily used to determine the Persona's short-term affective state - expressed through the system's affective output channels as gestures and speech inflection. However, emotions are also able to directly influence the choice of phrase within the conversational dialogue, and even dynamically adjust (within a pre-set range) the Persona's personality trait values. Likewise the Persona's personality traits: (a) help to bias the motivational profile of the agent - and thus determine the importance of the agent's high-level goals (to which events are compared during the emotion generation process); and (b) steer the conversational dialogue goals - extravert characters are more likely to initiate small-talk.

The Persona's *affective reasoning* process is based on the "Emotion Process" described in [7] - we only attempt to model the effects emotional states have on the behaviour of the agent, and do not claim to actually generate emotions within the architecture itself (i.e. there is no analogue of interruption of attentive processing). The *Information Coding* and *Primary Appraisal* components classify incoming *action events* and appraise them with respect to agent concerns. After classification, filtered events are passed to the *Output Generator* module as response requests through the two parallel affective information processing channels. The reactive module handles the Persona's immediate reactions to user or system events, whereas the deliberative module produces more controlled reactions (in particular, it is responsible for determining the contents and the structure of the dialogue contributions).

The rendering of the Persona's actions is performed by the *Output Generator* module, which generates and co-ordinates speech, facial expressions and body gestures. To allow the output generator to take into account the Persona's affective state, the deliberative and reactive modules annotate the actions to be executed with appropriate mark-ups (in addition to personality traits and emotion label terms - i.e. happy, sad, fear - we will also use the general emotional dimensions of *Valence* and *Arousal*).

To model the Persona's personality, we use the broad social dimensions of *extraversion*, *agreeableness*, and *neuroticism*. We will initially attempt to model the Persona as an extravert, agreeable and emotionally-balanced character - [12] argues that people tend to prefer others based on the match and mismatch to their own personality (even though the exact relationship is still unclear and empirical studies have led to conflicting results). Among other things, this means that the Persona will tend to take the initiative in a dialogue, will be co-operative and will remain patient if the user asks the same question over and over again (although the later case could indicate that the Persona is failing in her goal to be helpful).

Within our architecture, we currently focus on goal-based emotions - whereby events are evaluated with respect to their desirability for the user's and/or the Persona's goals. We will also attempt to infer the user's affective state and use it to create a more sympathetic character. The user's affective state may be derived directly from the syntactic and semantic form (use of affective phrases) and the acoustic realisation (talking speed, volume etc.) of his or her utterances. Furthermore, we infer it indirectly by monitoring system deficiencies, such as errors of the speech recognition component or the inaccessibility of information servers. If such events occur, the Persona will try to positively influence the user's affective state by her behaviour, e.g. by suggesting alternatives or simply by showing sympathy.

6 Conclusions

The three projects described in this paper use personality and emotions to emphasise different aspects of the affective agent-user interface: (a) *Puppet* uses affect to teach children how the different emotional states can change or modify a character's behaviour and how physical and verbal actions in social interactions can induce emotions in others; (b) the *Inhabited Market Place* uses affect to tailor the roles of actors in a virtual market place; and (c) *Presence* uses affect to enhance the believability of a virtual character and produce a more natural conversational manner.

Although all three projects use the same basic psychological models for personality and emotions, they differ in the types of implementation strategies adopted - reflected by the different requirements of their respective application domains. In *Puppet* and the *Inhabited Market Place*, we explicitly hand code personality style and affective responses centrally, within the presentation plans themselves. This allows us to rapidly prototype our presentations, taking advantage of a domain that is more or less self-contained. This approach seems appropriate for applications which address the generation of highly stereotypic response patterns for simpler emotion types. For example, in the *Puppet* project, we want to send clear and reproducible affective

signals to the children who play with the system. In Presence, we made the conscious decision to clearly separate the affect modelling process from the more general purpose presentation planning process. This allows us to maintain a fairly complex, and yet consistent, model of our agent's affective state over a number of very different dialogue topics and application domains - and still produce an engaging variety of responses at the output stage.

As this broad range of application areas demonstrate, affect has an important role to play in user-agent interactions. However, affective user interfaces are still in their infancy, and much work is still needed to bring all the pieces of the jigsaw together. To use *affect* effectively, it must be used both at an appropriate level for the application domain, and as an all-encompassing component of the system - from graphic design to system architecture to application content.

Acknowledgements

The Puppet Project is funded by the European Community within the i3-ese programme (Experimental School Environments) and started in October 1998. Our project partners are: the University of Aalborg, Denmark, (Laboratory of Image Analysis), the University of Aarhus, Denmark, (Institute of Dramaturgy) and the University of Sussex, UK, (School of Cognitive and Computer Science). We would like to thank Bo Cordes Petersen from the Laboratory of Image Analysis at the University of Aalborg for providing us with the screenshot of the 3D environment.

Presence is the first project to be founded and financed internally by the DFKI. Presence is an internal co-operation between the Deduction and Multi-Agent Systems Groups and the Intelligent User Interfaces Group.

The Inhabited Market Place is funded by the BMBF (Bundesministerium für Bildung und Forschung).

References

1. André E. (1999). Applied Artificial Intelligence Journal, Special Double Issue on Animated Interface Agents, Vol. 13, No. 4-5.
2. André, E., Rist, T. and Müller, J. (1999). Employing AI Methods to Control the Behavior of Animated Interface Agents. Applied Artificial Intelligence 13:415-448.
3. Ball, G. and Breese, J. (1998). Emotion and Personality in a Conversational Character. Workshop on Embodied Conversational Characters. Oct. 12-15, Tahoe City, CA, pp. 83-84 and 119-121.
4. Damasio, A. R. (1996). Descartes' Error. London: Papermac. (first published 1994, New York: G. P. Putman's Sons.)
5. Ekman, P. (1992), An Argument for Basic Emotions, In: Stein, N.L., and Oatley, K. Eds. Basic Emotions, p. 169-200, Hove, UK: Lawrence Erlbaum.

6. Elliott, C. 92 (1992). The Affective Reasoner: A process model of emotions in a multi-agent system. PhD Thesis, Northwestern University, Institute for the Learning Sciences Tech. Report #32.
7. Frijda, N. H. (1986). The Emotions. Cambridge: Cambridge University Press.
8. Furnham, A. (1990). Language and Personality. In: H. Giles and W.P. Robinson (Eds.) Handbook of Language and Social Psychology. (pp. 73-95). Chichester, England UK: John Wiley & Sons.
9. Gallaher, P.E. (1992). Individual Differences in Nonverbal Behavior: Dimensions of Style. *Journal of Personality and Social Psychology*, 63(1): 133-145.
10. Hayes-Roth, B. and van Gent, R. (1997). Story-Making with Improvisational Puppets, Agent '97, Marina del Rey, CA, USA.
11. Höök, K., M. Sjölander, A.-L. Ereback, and P. Persson. (1999). Dealing with the lurking Lutheran view on interfaces: Evaluation of the Agneta and Frida System. In *Proceedings of the i3 Spring Days Workshop on Behavior Planning for Lifelike Characters and Avatars*. 125-136. Sitges, Spain.
12. Isbister, K. and Nass, C. (1998). Personality in Conversational Characters: Building Better Digital Interaction Partners Using Knowledge About Human Personality Preferences and Perceptions. Workshop on Embodied Conversational Characters. Oct. 12-15, Tahoe City, CA, pp. 103-111.
13. Lang, P. (1995). The emotion probe: Studies of motivation and attention. *American Psychologist* 50(5):372-385.
14. Lester, J. C., S. G. Towns, C. Callaway, J. L. Voerman and P. J. FitzGerald. (1999). Deictic and emotive communication in animated pedagogical agents. In: Cassell et al. *Embodied Conversational Agents*, MIT Press, to appear.
15. Machado, I. and Paiva, A. (1999), Heroes, Villains, Magicians,...: Believable Characters in a Story Creation Environment, AI-ED '99 Workshop on Animated and Personified Pedagogical Agents, Le Mans, France.
16. MacIlhagga, M. and George, P. (1999). Communicating Meaningful Emotional Information in a Virtual World. In: *Proc. of the Workshop on Affect in Interactions - Towards a new Generation of Interfaces*, held in conjunction with the 3rd i3 Annual Conference, Siena, Italy, pp. 150-155.
17. Martinho, C. and Paiva, A. (1999). Pathematic Agents: Rapid Development of Believable Emotional Agents in Intelligent Virtual Environments. In: *Proc. of the Third Annual Conference on Autonomous Agents*, Seattle, WA, pp. 1-8.
18. McCrae, R. R., and John, O. P. (1992). An introduction to the five-factor model and its applications. Special Issue: The five-factor model: Issues and applications. *Journal of Personality* 60: 175-215, 1992.
19. Microsoft Agent: Software Development Kit, Microsoft Press, Redmond Washington, 1999. URL: <http://microsoft.public.msagent>
20. Ortony, A., Clore, G. L., and Collins, A. (1988). The Cognitive Structure of Emotions. Cambridge: Cambridge University Press.
21. Picard, R. W. (1997). Affective Computing. Cambridge, Mass: The MIT Press.
22. Reichenbach, I., and Masters, J. (1983), Children's use of expressive and contextual cues in judgements of emotions. *Child Development*, 54, p. 102-141.
23. Reilly, W. S. (1996). Believable Social and Emotional Agents. Ph.D. Thesis. Technical Report CMU-CS-96-138, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA. May 1996.

24. Scaife, M., Rogers, Y. (1996), External Cognition: How Do Graphical Representations Work? *International Journal of Human-Computer Studies*, 45, p. 185-213.
25. Scherer, K. R (1979). Personality Markers in Speech. In: K.R. Scherer & H. Giles (eds.). *Social Markers in Speech*, Cambridge University Press: Cambridge, pp. 147-209.
26. Simon, H. (1967). Motivational and emotional controls of cognition. Reprinted in *Models of Thoughts*, Yale University Press, (1979), pages 29-38.
27. Sloman, A. (1999). Architectural Requirements for Human-like Agents Both Natural and Artificial. (What sorts of machines can love?). To appear in K. Dautenhahn (Ed.) *Human Cognition And Social Agent Technology*, John Benjamins Publishing.
28. van Mulken, S., André, E. and Müller, J (1998). The Persona Effect: How Substantial is it? In: *Proc. of HCI'98*, Sheffield, pp. 53-66.
29. Velásquez J. D. (1999). *Proc. of Agents '99 Workshop on Emotional Architectures*, Seattle, WA.

Why Should Agents Be Emotional for Entertaining Users? A Critical Analysis

Paola Rizzo*

IP-CNR – Institute of Psychology, National Research Council of Italy
Viale Marx 15, I-00137 Rome, Italy
paola@ip.rm.cnr.it
<http://pscs2.irmkant.rm.cnr.it/users/paola/html/home.html>

Abstract. Emotional agents can play a very important role in all those computer applications where the user does not (only) want to perform a task, but (also) to have a more engaging experience than with traditional systems. In fact, it is commonly assumed that the agents' ability to process and display affective states, and to show emotional reactions, is crucial for causing a more enjoyable interaction between agent and user. This work proposes to analyze such assumption from a critical viewpoint, identifying several open issues that are worth debating in the community and studying through empirical methods.

1 Introduction

User interfaces are becoming more and more friendly, adaptive, and sensitive to the user's needs, preferences and wants. Along with this trend, computers are starting to comply with the users' wish to be pleasurably engaged while or besides performing some task, which may be particularly important in the case of children, young people, and unexperienced users. The issue of user's entertainment is to be "taken seriously", since it can deeply influence the effect of all kinds of software applications, and even make the difference between a successful and an unsuccessful interface. In fact, it has been observed that two main factors motivate the use of computer technology: the perceived usefulness and the perceived enjoyment [16,7]. While the former acts as an "extrinsic" motivator, in that users may perceive a software application to be instrumental for obtaining valued outcomes, the latter is an "intrinsic" motivator, because it causes people to use an application for no reason other than the process of using it *per se*. Indeed, computer-based entertainment can not only be an interesting purpose in itself (as in the case of PC games, multi-user virtual environments, or software toys), but can also increase the user's satisfaction, boost her motivation to interact with an application (which, for example, is very important in the case of "edutainment"), help relieve the tensions that are usually related to the

* The author is currently supported by a scholarship from the Committee 12 (Information Science and Technology) of the National Research Council of Italy. She is extremely grateful to Maria Miceli for her precious comments.

performance of difficult tasks, and give the opportunity to have a break from hard work. Such properties of entertainment come from its basically emotional nature and impact.

One way interfaces can be made more enjoyable is by means of “believable agents” (also known as “lifelike computer characters”, “synthetic agents”, “virtual actors”, and “animate characters”): these computational systems are built with the purpose of provoking in the user the “illusion of life” or, in other words, the impression of interacting with a “living” virtual being that has its own feelings, desires, and beliefs.¹ For example, instructional software systems can be enhanced by lively anthropomorphic or animal-like tutors [22,34]; multimedia presentation systems might look more likeable if they present material through expressive animations [2,1]; PC desktops may be more exciting if populated by “virtual pets” that interact with the user and with one another in interesting ways [10].

A strong argument in favor of using believable agents in interfaces is that they might make interacting with computers much easier and nicer, enable (especially naive) users to adopt communicative styles similar to those typical of human-human communication [2], and increase the level of interactivity and socio-emotional engagement produced by traditional applications. It is also commonly believed that the agents’ ability to process and display affective states, and to show emotional reactions, is crucial for improving the agents’ believability, for eliciting emotions in the user, and consequently for causing a more entertaining interaction between agent and user [31,3]. Therefore, several research projects are aimed at realizing *emotional agents* (EAs) (e.g. [9,33,24]).

It is important to remark that the agents’ believability and emotional behavior are by no means just a matter of visual appearance. In fact, an agent may be perceived as anthropomorphic even when it is not graphically portrayed in a human- or animal-like fashion [38], or when it is not visually represented at all, as in the case of chatterbots [25]. What is most important in causing the “illusion of life” is the agent’s *behavior*, rather than its look. Analogously, the agent’s emotional manifestation is not just confined to the physical exhibition of facial, vocal, or bodily expressions, but it involves the whole agent’s behavior. For instance, an agent that feels angry at somebody might try to take revenge, rather than displaying a peculiar facial or bodily expression.

There are some arguments against the usability of agents: it is suspected that they might generate wrong expectations about their own capabilities, compromise the user’s feeling of control over the tasks being performed, and reduce her sense of responsibility towards the accomplishment of the tasks [28,40]. However, these criticisms are usually addressed to one particular class of agents, namely those which reduce the user’s workload by acting on her behalf, for instance by filtering incoming e-mail messages or selecting web pages according to her preferences. Such agents are usually contrasted with other kinds of HCI metaphors, e.g. direct manipulation, that give greater control to users [40]. The focus of this

¹ This notion of believability is different from the perceived “credibility” of computer applications, which refers to the delivery of trustworthy information. [44]

paper is not on agents for delegation purposes, but rather on personified agents that can be conceived as sorts of partners or companions to their users, and that have as one of their major roles to entertain them, especially by manifesting emotions. Although these agents are not totally free from the criticisms mentioned above, the ultimate answer as to whether they are effective and usable will come from empirical research.

EAs are a promising direction of research on entertaining interfaces, but their realization and use seem to rely on a set of assumptions that have not yet been made explicit and studied in great depth. This work attempts to identify, analyze, and possibly criticize such assumptions. In our view, this analysis can be useful in two ways. On the one hand, it can help understand what are the main problems to address when realizing EAs. On the other hand, it can favor some empirical validation of the most common hypotheses regarding the relationships between entertainment, emotions, and believability, in order to build more effective agent-based interfaces. In general, the analysis is intended as a stimulus and a contribution to a hopefully wide discussion among researchers working in the EAs area.

1.1 A Note About the Concept of Entertainment

Generally speaking, entertainment can be defined as something that produces a state of mind, mostly emotional in nature, in which we feel engaged, usually with a sense of interest, pleasure, or enjoyment. There are many kinds of objects, events, or activities that can cause this state and which one refers to as different forms of entertainment: music, arts, games, literature, etc. Entertainment is not necessarily limited to positive emotions, such as fun or amusement. Think for example of the gloom we may experience when looking at some paintings by Edvard Munch, or of the “pleasant” fear we feel when watching a scaring movie: though negative, these feelings can be engaging and self-motivating for us, so we may actively look for them.

Since entertainment causes engagement, there might be a risk involved in making agent-based interfaces entertaining, because they could distract the user from the task she has to perform. However, such a view is partially misleading. In fact, the purpose of an entertaining interface is not to divert the user from her job, but rather to let her become more involved in it. As suggested by Laurel, “The user’s goals for a given application may be recreational, utilitarian, or some combination of both, but it is only through *engagement* at the level of the interface that those goals can be met” ([18] p. 69, original emphasis). Of course, the appearance, behavior, and role of an EA within an interface must be carefully designed, in order to avoid detrimental effects on the user’s performance, and to favor her involvement in the task (see for instance the criteria used in [22] for appropriately selecting the agent’s behavior and look according to pedagogical aims). In other words, the EAs’ entertainment potential should be aimed at supporting the user’s intrinsic motivation to using a software application, and this should be functional to the extrinsic motivation of using the application for achieving some desired result.

2 Emotional Agents: What Are the Underlying Hypotheses?

Since EAs are primarily built for entertaining users, one could wonder which properties of these agents make them entertaining, and which agents' features support those properties. Looking at the relevant literature, it seems that these important questions are answered by some typical presuppositions or tacit hypotheses, that influence how EAs are realized and used in computer interfaces. In our view, the hypotheses can be made explicit as follows:

Hypothesis 1: entertainment is a function of the character's ability to cause the "illusion of life" (believability);

Hypothesis 2: entertainment is a function of the character's ability to induce emotions;

Hypothesis 3: believability crucially depends on the agent's ability to show emotional behaviors;

Hypothesis 4: believability also depends on the manifestation of a marked personality, which in turn is based on an individual style of emotional behaviors;

Hypothesis 5: the user's emotional responses to an agent depend on the latter's emotional behavior;

Hypothesis 6: the agent's ability to show emotional reactions depends on its ability to represent and process emotional states.

The hypotheses appear to be related to each other according to an instrumental hierarchy, as shown in figure 1, that can also be read in a bottom-up direction: agent designers usually assume that the ability to represent and process affective states (which can be brought about by events affecting an agent's goals, preferences, or values) enables the agent to display emotional reactions; these in turn are deemed essential for enhancing the agent's believability, for characterizing its unique personality, which also contributes to believability, and for arousing emotions in the observer, possibly through an empathic process; finally, the agent's believability, and its ability to induce affective reactions in the user, are expected to be entertaining for the latter. As already mentioned, the user's entertainment can be viewed either as an end in itself, or as functional to other affective states, such as satisfaction, motivation, and the like, that can influence the effectiveness of several kinds of computer applications.

3 Analyzing the Hypotheses

Let us now try to trace each hypothesis back to the relevant literature and explain it, possibly looking for some counterexamples. The following questions drive our analysis:

- are we really sure that emotion manifestation and processing are crucial for making agents entertaining? If so, to what extent?

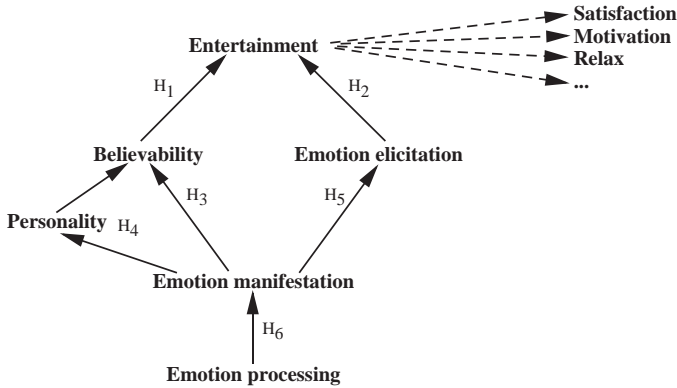


Fig. 1. Emotional agents: Underlying basic hypotheses

- are there alternative features of the agents, or of the interaction between agents and users, that could explain the agents’ entertaining ability?

Before starting our analysis it is to be remarked that this work does not address the issue of which properties of an agent’s architecture would make it really able to “have” or to “feel” emotions [31,6], nor why emotions could be deemed essential in a complete autonomous agent [41]. Rather, our focus is on those emotion-related characteristics of agents that are supposed to make them entertaining.

3.1 H₁ and H₂: Believability and Emotion Elicitation Are Functional to Entertainment

These two hypotheses are dealt with together because, as we will try to show, they are closely related to one another. They both give an initial general answer to the question of which agent’s characteristics are expected to cause the user’s entertainment.

The kind of entertainment EAs are most close to is *drama*, that basically consists in the representation of a play, performed by actors who impersonate different characters. Here we refer to a very broad concept of dramatic representation, that could be performed in several forms and through different media (e.g. cinema, TV, radio), and not just by human actors (think for instance of animation movies and puppets).

When is a character successful in entertaining the audience? Two abilities seem to be quite important. The first is the ability to let the spectators suspend their disbelief, that is to give them the illusion that what they are watching is reality. Laurel ([19] p. 113) tells us that the “willing suspension of disbelief”, concept introduced by the critic and poet Samuel Taylor Coleridge, “is the state of mind that we must attain in order to enjoy a representation of action. [...] we must temporarily suspend (or attenuate) our knowledge that it is ‘pretend’. We

do this ‘willingly’ in order to experience other emotional responses as a result of viewing the action”. This also relates to the second crucial ability a character should have, namely that of eliciting emotions in the audience. Usually this is obtained by the character through the manifestation of various emotional states (although later on we will try to show that this is not the only way of causing emotions in the observers).

It is interesting to notice that these two abilities can support each other: the more a character succeeds in producing the illusion of life, the more chances it will have to elicit emotions in the audience, and consequently a better illusion will be produced. However, despite the strong connection between the two abilities, it is to be remarked that emotion elicitation can be produced by an actor independent of the illusion of life. This happens in those performances, different from traditional drama, that are not strictly based on a representation of action: think for instance of an actor who tells jokes to the audience, or who reads poems or stories. In general, emotion elicitation is directly functional to any kind of entertainment, such as music or visual arts.

EAs are often considered similar to traditional characters: if well designed, they are expected to produce an “illusion of life” and enable the audience to experience strong vicarious feelings. Agents can be conceived as “actors” that play the role of a character assigned to them by the designer or “author”. Actors play their roles in interaction with an audience (the user) and might try to follow a plot that involves other actors and the audience itself. Since the latter’s behavior is unpredictable, there may be a “drama manager” that gives hidden directions to the actors so as to adjust the plot according to the occurring situations, in the attempt to create a dramatically rich experience for the human users [23,13,39].

These first two hypotheses, although difficult to verify empirically, are hardly questionable; they can actually be considered as basic assumptions. In fact, we have many examples of their validity from drama: the more believable a character is, the more likeable it is, and the emotions it can induce greatly contribute to the spectator’s entertainment.

3.2 H₃: Believability Crucially Depends on the Agent’s Ability to Show Emotional Behaviors

Assuming that the user’s entertainment is caused by the agent’s believability, one should wonder what makes an agent believable. In the case of a human character, the suspension of disbelief would depend both on her acting skills and on the story plot; but what about a synthetic character? In this case an agent, before being considered as a (good or bad) character, should first of all be perceived as a human-like being by the audience.

Intuitively, this could be obtained by endowing the agent with as many anthropomorphic features as possible. However, the “anthropomorphization” of an agent, analogously to what is done with animated characters, does not necessarily consists in giving it a realistic human or animal appearance, but rather in choosing for it the right set of behavioral features that would drive the observers towards naive psychological *attributions*. Think for example of Pixar’s

short film *Luxo Jr.*: the two animated characters are portrayed as lamps, without face, arms, or legs; however, it is very natural for us to perceive them as mother and child, to recognize the typical playful attitude of the latter or the scolding attitude of the former and, in general, to interpret their behavior in an anthropomorphic way, with a very engaging and entertaining effect.

A good analysis of requirements for believability is provided by Loyall [23]. His analysis is mostly inspired by the work of two Disney animators [42], who have a remarkable experience in producing the illusion of life in animated characters. Their work has been integrated by Loyall with sources from playwriting and drama, and filtered by his experience in building computational believable agents. Loyall's list of requirements, that appears to be more or less shared by the community working on believable agents, includes personality, emotion, self-motivation, and social relationships.

Though many features seem to contribute to believability, it is quite evident that a strong emphasis in the EAs community has been given to emotions (e.g. [33,9,21]). It is generally assumed that the better emotions are manifested by an agent through its appearance and behavior, the more believable it is. According to Joseph Bates, pioneer and leading researcher in this area, "The emotionless character is lifeless. It is a machine." ([3] p. 123).

However, some counterexamples can be found that could weaken this hypothesis. A famous case of emotionless character that appears very believable and engaging is ELIZA, a system of natural language understanding developed by Joseph Weizenbaum in the 60's [45], based on pattern-matching techniques and on a script, i.e. "a set of rules rather like those that might be given to an actor who is to use them to improvise around a certain theme" ([46] p. 3). The most famous conversational role played by ELIZA is that of DOCTOR, an hypothetical Rogerian psychotherapist engaged in an initial interview with the patient. Weizenbaum has interestingly noticed that "I was startled to see how quickly and how very deeply people conversing with DOCTOR became *emotionally involved* with the computer and how unequivocally they *anthropomorphized* it" and that even "short exposures to a relatively simple computer program could induce powerful delusional thinking in quite normal people" ([46], pp. 6-7, added emphasis). It is interesting to notice that ELIZA, despite its lack of emotion manifestation, is not only very believable, but also able to elicit emotions in the users, which seems also to contradict H₅ (see below). Actually, ELIZA is a very good example of a computer program that can be perceived as human-like by the users, without needing realistic anthropomorphic representation or behaviors; the "secret" of its success seems to be based on the desires, principles, and beliefs users attribute to it, according to how the interaction goes on.

In general, it is to be remarked that anthropomorphism is a very general, strong, and spontaneous tendency that can be elicited by many kinds of non-human behavior and characteristics. For example, in a classical psychological study by Heider and Simmel [15], it has been observed that people who are shown a short animated film of geometric shapes moving around, interpret it as a story in which first a big triangle chases a little one, then the latter locks

the former into a house (a rectangle with a moving-flap “door”), and finally the house explodes. Subjects tend even to attribute a gender to the moving forms. In a similar vein, Nass and co-authors have shown that “a *minimal* set of characteristics associated with humans provides sufficient cues to encourage users to exhibit behaviors and make attributions toward computers that are considered appropriate only when directed at other humans” ([27] p. 111, original emphasis). This has been explained by taking into account the *social* nature of interactions between humans and computers or other new media. [32]

3.3 H₄: Believability Crucially Depends on the Manifestation of an Emotion-Based Personality

Another feature that is usually deemed very important for believability is the manifestation of a marked personality: in fact, an agent that displays individual styles of behavior is supposed to be more lifelike and appealing to the user, and better distinguishable from other agents. Therefore, another major problem faced by many researchers is how to model personality in artificial agents. [43]

Given the importance attributed to emotions, personalities are often realized by letting the agents display individual differences in emotional behaviors. The cognitive theory of emotions by Ortony, Clore and Collins [29] usually provides a useful basis for this kind of approach. According to these authors, as well as many others (e.g. [20]), the elicitation and differentiation of emotions occur through an appraisal of situations or events that are particularly important to a person. In other words, the person evaluates how each situation she faces affects her own concerns, goals, values and the like, and the results of the assessment (in terms for example of unexpectedness of an event, or desirability and likelihood of an outcome) can elicit several distinct emotions and their corresponding behavioral expression. By means of suitable approximations and adjustments, this theory has been used by several agent designers (e.g. [4,9,33,24]) for modeling personalities by varying the mapping between internal representations, emotions and behaviors. Each emotion that is triggered by the evaluation of the current state of a value, goal, or preferred object (e.g. the success, failure, or threat to a goal) can be associated with a class of reactions typical of each personality: for example, fear can be associated with a class of escape reactions for realizing a fearful personality, or with fight behaviors in order to realize an aggressive personality.

The two-faced hypothesis that personality is crucial for believability, and that it should be based on individual styles of emotional behaviors, can be subject to a couple of criticisms.

The first is that personality, as well as emotion manifestation, is probably not so crucial for creating the illusion of life. Some agents that are not endowed with any personality can appear quite believable: think for instance of the already mentioned ELIZA program, of agents such as Steve [34], or even of robotic agents (e.g. [30]). As in the case of emotion manifestation, one could conclude that the tendency towards anthropomorphism is so strong that it does not necessarily

need to be supported by the agent's manifestation of personality, although the latter favors and reinforces such tendency.

The second criticism that can be aroused concerns the relationship between personality and emotions. It is true that real people can often be characterized one with respect to another according to their affective tendencies, and this is particularly significant in the case of dramatic characters (think for instance of the classical distinction between “good” and “evil” characters). This argument provides a strong support to emotion-based models of personality for agents. Nevertheless, alternative models of personality can be used (e.g. [11,37]), which seem to be at least as much expressive.

As an example, personality can be modeled through a “goal-based” approach [5,35,36], according to which each agent's personality is defined as (1) a cluster of “General Goals”, or G-GOALS, having different priorities and (2) a set of goal-based preferences over actions and plans. On the one hand, G-GOALS represent abstract states or events that an agent recurrently attempts to achieve and maintain through its behavior. Each G-GOAL represented in the personality cluster is assigned a different priority, that specifies its importance with respect to other G-GOALS in the cluster: this characterizes both each agent's coherent behavioral organization and the individual differences among agents. On the other hand, preferences are determined by considering the side-effects of actions on the G-GOALS, i.e. changes in the world that affect goals which are not actively pursued when a given plan is executed. If the affected goals are important in the personality goal cluster, problems or opportunities may arise: for instance, if a G-GOAL can be opportunistically achieved by performing an action having a positive side-effect on that G-GOAL, that action is preferred over “neutral” actions, while an action having a negative side-effect on an important G-GOAL will be rejected.

The choice of a goal-based model of personality could be motivated by psychological findings about how people make attributions with regard to others' behavior. In fact it has been observed that, when making attributions, people show a strong tendency to prefer personal factors, while neglecting or undervaluing environmental causes [14]. According to Jones and Davis [17], people also tend to produce “corresponding inferences”, that is to infer stable dispositions and characteristics from the intentions that are attributed to a person's behavior. For example, if someone performs a hostile behavior, the “corresponding” inference is that the behavior is determined by an intention of attacking, which in turn is caused by an aggressive personality. Since people naturally tend to attribute intentions and corresponding dispositions to others' behavior, it would seem reasonable to endow agents with specific goals that strongly characterize and differentiate them, and to make these goals manifest in the agents' behavior, so as to support and strengthen the dispositional bias.

3.4 H₅: Emotion Elicitation Crucially Depends on the Agent's Manifestation of Emotions

As already mentioned when discussing the first two hypotheses, in the context of drama, an actor's ability to bring out affective involvement in the audience is supposed to be based on his skills in expressing emotions. In the context of synthetic personas this view is well stated by Bates: "If the character does not react emotionally to events, if it doesn't care, then neither will we." ([3] p. 123). Such hypothesis seems to be hardly questionable: in fact, there are countless examples of characters whose success comes from their emotional expressiveness. This seems to be grounded on an empathic process: when watching a character, people tend to identify with it, and to feel the same emotions it shows. However, the matter seems to be much more complex than this: at least four issues are worth discussing.

The first is that the situation in which a character is involved can elicit emotions by itself, even if the character does not display any affective reaction. Think for instance of those classic comic scenes where a cake is thrown at someone's face, or where somebody falls over a banana-peel: we are amused independently of the character's emotional expression. An analogous argument can be done relative to those situations that trigger negative emotions in the audience independently of the perception of affective states in the involved characters. Understanding the reasons why we are entertained by some situations is beyond the scope of this work;² it is just worth remarking that emotion manifestation per se is not crucial for eliciting emotions and consequently some kind of entertainment.

A second observation is that, in the situations that automatically elicit a typical emotion in the audience, the character may actually display some affective reaction that will not necessarily cause an empathic response in the viewers. For instance, a character who has just got a cake in the face may show to be angry or sad, but the observer will feel a rather opposite emotion. As a matter of fact, the contrast between the funny situation and the character's negative expression may be even more entertaining than no emotion manifestation in the latter. This points to the argument that empathy is not the only kind of relationship that can exist between a character's and a viewer's emotions.

A third issue is that emotion elicitation in the audience can occur not only through the perception of a character's emotions, but also through the *attribution* of affective states to the character. The attribution is based on an understanding of the relationship between a given situation and the supposed character's goals, attitudes, or values. As a typical example, if we know or believe that the hero of a movie is in love with the heroine, and an evil character threatens her, we will assume that the hero is angry, and we may feel angry ourselves, even though the expected emotion is not displayed by the hero. A remarkable example of this phenomenon is given by Buster Keaton, the famous comic actor of silent cinema. Keaton's acting style is very peculiar: no matter what happens to him,

² For a review of the psychological theories of humour see for instance [12].

he is always “wearing” a kind of sad face. Nevertheless, it is quite natural for us to attribute to him various affective states according to the situations he encounters.

A final issue to be tackled concerns the emotions that are caused by interacting with characters in a first-person perspective. The problems we have addressed so far are based on a third-person, non-interactive view of the relationship between a persona and an observer, where the latter passively watches the performance of the former, possibly identifying and “empathizing” with it or feeling dissimilar emotions. This is a simplified picture of how agents might elicit emotions in the user. Actually, what makes agents different from traditional characters is their ability to engage the user in an interactive experience, where she actively participates in the ongoing action. In this case, the user’s emotional reactions can be affected by how *her own* goals and beliefs come to interact with those of an agent in a given situation. Think for example of “enemies” in computer games: the user may feel scared when threatened by them, and relieved when she kills or escapes from them. Once again it is clear that the user’s feeling is not necessarily determined by the affective outlook of the agents (consider for instance the emotionless enemies of the popular *Pacman* game), although it can be further supported and reinforced by suitable emotional expressions in the agents.

Of course, this discussion of the hypothesis that emotion elicitation in the users depends on emotion manifestation by the agents has touched several controversial issues, each of which would deserve a separate study. However, the aim of the discussion is not to provide ultimate explanations of those issues, but rather to point out that the hypothesis needs to be considered with great care, if one wants to build agents that are effective in entertaining users through emotion elicitation.

3.5 H₆: Emotion Manifestation Crucially Depends on Emotion Processing

The final hypothesis we will analyze states that agents should be able to represent and process emotional states in order to automatically generate emotional behaviors. Since it is unquestionable that an agent could display emotions without having them (just to take a trivial example, think of a simple smiling face on the computer screen), why should a designer want to endow her agents with affective processing mechanisms? Because they make agents able to display a broader range of more complex emotions. As stated by Reilly, “the agent will be emotionally richer if there are more things [such as goals, desires, and analogous cognitive structures] to have emotions about and more ways to express them.” ([33], p. iii).

In principle, this argument is not disputable at all. Fast, primary emotions such as fear or anger in face of danger, that are usually associated to elementary behaviors like escape or aggression, could be caused in the agents by means of very simple condition-action rules. By contrast, more complex affective behaviors are to be supported by some cognitive mechanism. For instance, there is no

way to bring out an emotion like “hope”, and its corresponding behavioral expressions, without representing the goal an agent believes to be likely to succeed (i.e. the “object” of the hope) and its importance.

However, going back to H_5 , it is worth analyzing how the way complex emotions are processed may influence the way such emotions are manifested by agents, and how this in turn may affect the users’ entertainment. In fact, in the context of entertainment, one could assume that the broader the range of emotions manifested by agents, the more engaging is the interaction for the user.

Intuitively, this assumption seems to be correct. However, two controversial issues are worth considering. The first is whether users really or always need to perceive complex, cognitively generated emotions in order to be entertained. Some answers to this question could come from the preceding discussion of H_5 , where it has been pointed out that entertainment could be simply caused by the situations in which agents are placed, rather than by the agents’ emotional expressions, and that users could be entertained also by agents who display only a few basic emotions, as in the case of computer games opponents.

A conclusion one could draw from these observations is that agents should not necessarily be able to represent and reason about affective states: rather, they would just need to display a small number of fundamental emotions relative to a few significant classes of situations, together with some ambiguous behaviors that would remain open to subjective interpretation by the user. In fact, as often remarked in this work, people tend to make complex attributions even out of very elementary behaviors, which means that an agent’s emotional simplicity would be counterbalanced by the user’s interpretation process.

This is also related to the second issue to consider, concerning the inherent difficulty of driving the user’s interpretation of the agents’ behaviors towards a desired direction. As mentioned earlier, most models of emotion processing in agents are built on a psychological theory [29] of the human mechanisms that are triggered by events affecting internal representations, and that lead to emotion differentiation and manifestation. The theory is not aimed at explaining how a given emotional manifestation would be perceived and interpreted by an external observer. So, there might happen to be some mismatch between the agent’s emotion as it is manifested and the user’s interpretation. Although agent designers also use techniques derived from animation in order to better convey the agents’ internal processes and characteristics [23], the problem remains hard.

As interestingly suggested by Sengers [38] the focus of designers, rather than on the agents’ internal mechanisms, should be on the requirements for making agents externally comprehensible. In order to do this, she proposes to structure the agents’ behavior according to the “signs” and “signifiers” it is intended to communicate, coupled with a sign management system and a behavioral transition mechanism.

A detailed discussion of the possible solutions to the user interpretation problem is out of the scope of this work. Rather, our aim is to point out that complex emotion processing techniques may not produce the intended result, or may even hinder a successful manifestation of emotions for user entertainment.

4 Conclusion

We believe that EAs have a strong potential in entertaining users, and therefore should play an important role within new computer interfaces. However, as shown by the analysis above, the design and realization of EAs seem to be based on a set of hypotheses that are far from being completely understood. On the one hand, it is plausible that the EAs' entertaining potential comes from their ability to provoke the illusion of life and to elicit emotions in the users. On the other hand, it is not totally clear what supports these abilities, and which is the specific role played by emotions. In fact, it has been observed that emotion manifestation and processing are not necessarily crucial for creating the illusion of life, for eliciting emotions in the users, or for realizing a believable personality. Such arguments point out that the hypotheses above should not be taken for granted, but rather should be empirically tested.

Some experimental investigations about the effectiveness of animate characters (e.g. [26,21]; for a review see [8]) have focused mainly on two aspects: (a) some objective measure of the users' performance (like the number of solved problems, or the percentage of remembered items after their presentation), in order to see whether the performance is improved by the presence and/or help of an agent, and (b) several subjective measures of the characteristics (like intelligence, likeability, utility) that can be attributed by the users to the agent, and that are supposed to mediate the agent's effectiveness.

Unfortunately, these studies are difficult to compare with one another because of their different experimental settings, dependent variables, etc., and consequently it is also hard to derive from them some useful generalizations. But, more importantly, the empirical works carried out so far do not seem to be based on a causal model that could explain the possible interrelationships among variables (i.e. among the subjective and the objective measures). In our view, the hierarchy of hypotheses proposed in this work could be considered as a frame of reference where the possible cause-effect links between variables are explicitly represented; as such, it could help devise new experiments aimed, on the one hand, at testing those links one by one and, on the other hand, at verifying whether the model should be modified by the introduction of other variables and related links/hypotheses.

References

1. E. André, M. Klesen, P. Gebhard, S. Allen, and T. Rist. Integrating models of personality and emotions into lifelike characters. In A. Paiva and C. Martinho, editors, *Proceedings of the Workshop on Affect in Interactions of the i³ Annual Conference*, pages 136–149. Siena (Italy), October 21–22 1999. 167
2. E. André, T. Rist, and J. Müller. Integrating reactive and scripted behaviors in a life-like presentation agent. In *Proceedings of the Second International Conference on Autonomous Agents*, pages 261–268. ACM Press, New York (NY), 1998. 167
3. J. Bates. The role of emotion in believable agents. *Communications of the ACM*, 37(7):122–125, 1994. 167, 172, 175

4. J. Bates, A. B. Loyall, and W. S. Reilly. Integrating reactivity, goals, and emotion in a broad agent. In *Proceedings of the Fourteenth Annual Conference of the Cognitive Science Society*, Bloomington (IN), 1992. Also available as: Technical Report CMU-CS-92-142, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA. 173
5. J. Carbonell. Towards a process model of human personality traits. *Artificial Intelligence*, 15:49–74, 1980. 174
6. C. Castelfranchi. To believe and to feel: The case of “needs”. In D. Canamero, editor, *Proceedings of AAAI Fall Symposium “Emotional and Intelligent: The Tangled Knot of Cognition”*, pages 55–60. AAAI Press, Menlo Park (CA), 1998. 170
7. F. D. Davis, R. P. Bagozzi, and P. R. Warshaw. Extrinsic and intrinsic motivation to use computers in the workplace. *Journal of Applied Social Psychology*, 22:1111–1132, 1992. 166
8. D. M. Dehn and S. van Mulken. The impact of animated interface agents: A review of empirical research. *International Journal of Human-Computer Studies*, 1999. To appear. 178
9. C. Elliott. *The Affective Reasoner: A Process Model of Emotions in a Multi-Agent System*. PhD thesis, Northwestern University, Evanston (IL), May 1992. Technical Report no. 32. 167, 172, 173
10. A. Frank, A. Stern, and B. Resner. Socially intelligent virtual petz. In *Proceedings of the AAAI’97 Spring Symposium on “Socially Intelligent Agents”*. AAAI Technical Report FS-97-02, pages 43–45. AAAI Press, Menlo Park (CA), 1997. 167
11. A. Goldberg. Improv: A system for real-time animation of behavior-based interactive synthetic actors. In R. Trappl and P. Petta, editors, *Creating Personalities for Synthetic Actors*. Springer-Verlag, Berlin, 1997. 174
12. J. H. Goldstein and P. E. McGhee. *The Psychology of Humor*. Academic Press, New York (NY), 1972. 175
13. B. Hayes-Roth, L. Brownston, and R. van Gent. Multi-Agent collaboration in directed improvisation. In *First International Conference on Multi-Agent Systems*, pages 148–154. MIT Press, Cambridge (Mass.), 1995. 171
14. F. Heider. *The Psychology of Interpersonal Relations*. John Wiley & Sons, New York (NY), 1958. 174
15. F. Heider and M. Simmel. An experimental study of apparent behavior. *American Journal of Psychology*, 57:243–259, 1944. 172
16. M. Igbaria, S. J. Schiffman, and T. J. Wieckowski. The respective roles of perceived usefulness and perceived fun in the acceptance of microcomputer technology. *Behaviour and Information Technology*, 13(6):349–361, 1994. 166
17. E. E. Jones and K. E. Davis. From acts to dispositions: The attribution process in person perception. In L. Berkowitz, editor, *Advances in Experimental Social Psychology*, volume 2. Academic Press, 1965. 174
18. B. K. Laurel. Interface as mimesis. In D. A. Norman and S. W. Draper, editors, *User centered system design*, pages 67–85. Lawrence Erlbaum Associates, Hillsdale (NJ), 1986. 168
19. B. K. Laurel. *Computers as Theatre*. Addison-Wesley, Reading (MA), 1991. 170
20. R. S. Lazarus. *Emotion and adaptation*. Oxford University Press, New York (NY), 1991. 173
21. J. C. Lester, , S. A. Converse, S. E. Kahler, S. T. Barlow, B. A. Stone, and R. S. Bhogal. The persona effect: Affective impact of animated pedagogical agents. In S. Pemberton, editor, *Human factors in computing systems: CHI’97 Conference Proceedings*, pages 359–366. ACM Press, New York (NY), 1997. 172, 178

22. J. C. Lester and B. A. Stone. Increasing believability in animated pedagogical agents. In W. L. Johnson, editor, *Proceedings of the First International Conference on Autonomous Agents*, pages 16–21. ACM Press, New York (NY), 1997. 167, 168
23. B. Loyall. *Believable Agents: Building Interactive Personalities*. PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh (PA), 1997. Technical Report CMU-CS-97-123. 171, 172, 177
24. C. Martinho and A. Paiva. Pathematic agents: Rapid development of believable emotional agents in intelligent virtual environments. In *Proceedings of the Third International Conference on Autonomous Agents*, pages 1–8. ACM Press, New York (NY), 1999. 167, 173
25. M. L. Mauldin. Chatterbots, tinymuds, and the Turing test entering the Loebner prize competition. In *Proceedings of the 12th National Conference on Artificial Intelligence (AAAI'94)*. AAAI Press, Menlo Park (CA), 1994. 167
26. S. Van Mulken, E. André, and J. Müller. The Persona Effect: How Substantial Is It. In *Proceedings of HCI'98*, pages 53–66, Sheffield, UK, 1998. 178
27. C. Nass, J. Steuer, E. Tauber, and H. Reeder. Antropomorphism, agency, and ethopoeia: Computers as social actors. In *Interchi'93 Conference Proceedings. Conference on Human Factors in Computing Systems*, pages 111–112. ACM Press, New York (NY), 1993. 173
28. D. A. Norman. How might people interact with agents. *Communications of the ACM*, 37(7):68–71, 1994. 167
29. A. Ortony, G. Clore, and A. Collins. *The Cognitive Structure of Emotions*. Cambridge University Press, Cambridge (England), 1988. 173, 177
30. S. Penny. Embodied cultural agents: At the intersection of robotics, cognitive science and interactive arts. In *Proceedings of the AAAI'97 Spring Symposium on "Socially Intelligent Agents"*. AAAI Technical Report FS-97-02, pages 103–105. AAAI Press, Menlo Park (CA), 1997. 173
31. R. Picard. *Affective Computing*. The MIT Press, Cambridge (MA), 1997. 167, 170
32. B. Reeves and C. Nass. *The Media Equation: How People Treat Computers, Television and New Media like Real People and Places*. Cambridge University Press, Cambridge (MA), 1996. 173
33. W. S. Reilly. *Believable Social and Emotional Agents*. PhD thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, May 1996. Technical Report CMU-CS-96-138. 167, 172, 173, 176
34. J. Rickel and W. L. Johnson. Integrating pedagogical capabilities in a virtual environment agent. In W. L. Johnson, editor, *Proceedings of the First International Conference on Autonomous Agents*, pages 30–38. ACM Press, New York (NY), 1997. 167, 173
35. P. Rizzo. *Personalities in believable agents: A goal-based model and its realization with an integrated planning architecture*. PhD thesis, Center of Cognitive Science, University of Turin, Turin, Italy, 1998. 174
36. P. Rizzo, M. M. Veloso, M. Miceli, and A. Cesta. Goal-based personalities and social behaviors in believable agents. *Applied Artificial Intelligence*, 13(3):239–271, 1999. Special Issue on “Socially Intelligent Agents”, edited by K. Dautenhahn and C. Numaoka. 174
37. D. Rousseau and B. Hayes-Roth. Personality in synthetic agents. Technical report, Knowledge Systems Laboratory, Department of Computer Science, Stanford University, Stanford (CA), 1996. 174

38. P. Sengers. Designing comprehensible agents. In *Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI'99)*, pages 1227–1232. 1999. 167, 177
39. N. M. Sgouros, G. Papakonstantinou, and P. Tsanakas. A framework for plot control in interactive story systems. In *Proceedings of the 13th National Conference on Artificial Intelligence (AAAI'96)*, pages 162–167. AAAI Press, Menlo Park (CA), 1996. 171
40. B. Shneiderman and P. Maes. Direct manipulations vs interface agents: Excerpts from debates at IUI'97 and CHI'97. *Interactions*, 4(6):97–124, 1997. 167
41. A. Sloman and M. Croucher. Why robots will have emotions. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI'81)*, page 197. Morgan Kaufmann, San Mateo (CA), 1981. 170
42. F. Thomas and O. Johnston. *Disney Animation: The Illusion of Life*. Abbeville Press, New York (NY), 1981. 172
43. R. Trappl and P. Petta, editors. *Creating Personalities for Synthetic Actors*. Springer-Verlag, Berlin (Germany), 1997. 173
44. S. Tseng and B. J. Fogg. Credibility and computing technology. *Communications of the ACM*, 42(5):39–44, 1999. 167
45. J. Weizenbaum. ELIZA—A computer program for the study of natural language communication between man and machine. *Communications of the Association for Computing Machinery*, 9(1):36–45, 1965. 172
46. J. Weizenbaum. *Computer Power and Human Reason*. Freeman, San Francisco (CA), 1976. 172

Emotional Meaning and Expression in Animated Faces

Isabella Poggi¹ and Catherine Pelachaud²

¹ University of Rome Three, Department of Linguistics
Via del Castro Pretorio, 20, 00185 Rome Italy
`poggi@uniroma3.it`

² University of Rome “La Sapienza”, Department of Computer and System Science
Via Buonarroti, 12, 00185 Rome Italy
`cath@peano.dis.uniroma1.it`

Abstract. This paper shows that emotional information conveyed by facial expression is often contained not only in the expression of emotions per se, but also in other communicative signals, namely the performatives of communicative acts. An analysis is provided of the performatives of suggesting, warning, ordering, imploring, approving and praising, both on the side of their cognitive structure and on the side of their facial expression, and it is shown that the meaning and the expression of emotions like sadness, anger, worrying, uncertainty, happiness and surprise are contained in them. We also show that a common core of meaning is present in an emotion (surprise) as well as in other kinds of communicative signals (emphasis, back channel of doubt, adversative signals). We then argue on how the cognitive and expressive analyses of these communicative acts may be applied in the construction of expressive animated faces.

Keywords: Emotions, Conversational Agent, Performative, Embodiment, Communicative Act

1 Introduction

Research has shown that humans tend to treat computers as social characters [6,24]. As a consequence we may think that embodied artificial agents will help human-computer interaction. A recent study [11] proposes some guidelines on when and where to use such agents: education and entertainment are well appropriated fields. Agents may be helpful and may provide suggestions to users that have to make decisions in uncertain situations. Dialoguing with the agent to get more information, suggestion and criticism is often perceived positively by the user. Several researches [15,25,27] have suggested that the productivity and performance of a user is enhanced by the use of talking faces. The system is perceived as more engaging. In particular, showing emotional expression increases user attention; the user spends more time interacting with an agent with a stern face than one with a neutral expression [27].

In the construction of embodied agents capable of expressive and communicative behaviors, an important step is to reproduce affective and conversational facial expressions on synthetic faces [1,2,4,5,18,23]. The facial expression of affective states is the first and most obvious aspect of facial communication that has to be simulated in animated face; preceding works have shown that in humans emotions like fear, anger, happiness, sadness, surprise are expressed through specific facial muscular actions [13] and that these perceivable expression can be simulated in synthetic animated faces [3,16,19,26].

In this work, however, we want to show that emotional contents and their expression are not present only in strictly emotional facial expressions; some affective contents, as well their corresponding facial expressions, are also present in typically communicative signals that are not at first sight emotionally loaded, for instance in the expression of the performative of a Speech Act or, more generally, of any (verbal or nonverbal) Communicative Act. By presenting a compositional view of the facial expression of performatives, we analyse some performatives into their components both on the side of their cognitive structure and of their expressive muscular actions, and we show that on both sides some emotional components can be found. Then we explore how a typical facial expression of an emotion (the raising of the eyebrows) is also present in other communicative functions. We try to find out what is the common core of meaning that is shared by the facial expressions of all these functions. We conclude by showing how the creation of synthetic agents can use these results and how the computation of their behavior should be based on a semantic approach rather than on a surface approach.

2 Emotions in the Meaning of Performatives

The system we present here is based on a goal and belief model of action and communication [10,20]. In this model a communicative act corresponds to the minimal unit of communication. A communicative act is any verbal or nonverbal action through which a Sender has the goal that the Addressee get some beliefs about the Sender's goals and beliefs. To fully analyse a communicative act one has to consider two aspects: its signal and its meaning. The signal of a communicative act is a set of muscular activities or morphological changes that can be perceived by an Addressee. The meaning of a communicative act includes a performative and a propositional content, where the propositional content is the set of beliefs mentioned and the performative is the Sender's specific social-communicative goal, that is, what the Sender wants the Addressee to do in mentioning that specific propositional content. For example, in the sentence "I suggest you take your umbrella", 'suggest' is the performative and 'you take your umbrella' is the propositional content. In particular, in previous works on animated faces [21,22], we proposed a way to reproduce the facial expressions that convey the performative of a communicative act. In our model, the meaning side of a performative in a communicative act may be analyzed as a set of "cognitive units" [7], logical propositions whose predicates are primitive predicates

like Goal, Belief, and whose arguments are persons (Sender, Addressee), domain ‘objects’, domain ‘facts’, domain ‘actions’, emotions. The facial expression is represented discursively as perceivable states or movements of specific regions of the face (lips, eyebrows, cheeks), or more schematically in terms of Ekman and Friesen’s Action Units as defined in FACS (Facial Action Coding System) [14], that can be implemented in 3D facial model [19].

Suppose a Sender **S** suggests an Addressee **A** to do some action **a**. The meaning of the performative of suggesting (in its reading as a request, not as an information), that is the set of information that **S** wants **A** to believe when providing a suggestion, may be represented by the following cognitive units [21]:

- | | |
|--|---|
| 1. S has the goal that A do a | S requests A to do a |
| 2. S believes a is useful to a goal g of A | a is in the interest of A |
| 3. S believes 2. with low degree of certainty | degree of certainty |
- Cognitive Units of suggesting

On the side of the facial signal, the performative of suggesting is expressed by looking at the addressee, with the head a bit leaning forward and the eyebrows slightly raised. Sometimes, the global expression of the face corresponds to the complete cluster of cognitive units that makes up a performative; but in some cases we can find even a one-to-one correspondence among cognitive units and facial actions: for instance, in the case of suggestion head leaning forward may be a quite specific way to express that the requested action is not in the interest of **S** but of **A** (Cognitive Unit 2.); while a slight eyebrow raising may typically express perplexity or uncertainty (Cognitive Unit 3.). Focusing on the meaning of performatives, if we analyze a fair number of them, we can see that on the meaning level several performatives contain information about some affective state. Let us see some examples. The performative of imploring may be analyzed as follows:

- | | |
|--|---|
| 1. S has the goal that A do a | S requests A to do a |
| 2. S believes a is useful to a goal g of S | a is in the interest of S |
| 3. A has power over S | power relationship |
| 4. if A does not do a , then S will be sad. | potential affective state |
- Cognitive Units of imploring

The performative of imploration contains the affective state of sadness because, if I ask you something that is very important to me and that I cannot obtain without your help, I can anticipate that I will be sad if you do not fulfil my request. Now take the performative of a preemptory order:

- | | |
|--|---|
| 1. S has the goal that A do a | S requests A to do a |
| 2. S believes a is useful to a goal g of S | a is in the interest of S |
| 3. A has power over S | power relationship |
| 4. if A does not do a , then S will be angry. | potential affective state |
- Cognitive Units of preemptory order

Again, the performative of peremptory order contains the affective state of anger. If I request you to do something while assuming that I have power over you, I can show potential anger, since if you do not fulfil my request I'll be angry at you. Here is, moreover, the performative of warning:

- | | |
|---|---------------------------------------|
| 1. S has the goal that A believes c | S informs A of c |
| 2. S believes that not to believe c may cause
some goal of A to be thwarted | social relationship |
| 3. S has the goal that A 's goals are reached | |
| 4. S is worrying for A | potential affective state |
- Cognitive Units of warning

The performative of warning contains the affective state of worrying: warning means that I give you some information that is important for you, at the extent that if you did not know it something wrong could happen to you: thus, warning contains my being worried for you.

Let us now take the difference between approving and praising.

- | | |
|--|--|
| 1. S believes that A has done a | |
| 2. S believes that doing a is a good thing | S evaluates a |
| 3. S is happy | affective state |
| 4. 2. causes 3. | |
| 5. S has the goal the A believes 4. | S informs A of an evaluation |
- Cognitive Units of approving

- | | |
|---|--|
| 1. S believes that A has done a | |
| 2. S believes that doing a is a good thing | S evaluates a |
| 3. S is happy | affective state |
| 4. 2. causes 3. | |
| 5. S believes that A did a particularly well | S evaluates a |
| 6. S is surprise | affective state |
| 7. 5. causes 6. | |
| 8. S believes that A is good | S evaluates A |
| 9. 5. causes 8. | |
| 10. S has the goal that A believes 8. | S informs A of an evaluation |
- Cognitive Units of praising

In both approving and praising, **S** believes that **A** has done some action **a** and that doing **a** is a good thing, (see the Cognitive Units 1. and 2. in both performatives), and in both this makes **S** happy (Cognitive Units 3. and 4.). However, praising differs from approving in that **S** not only believes that doing **a** is good (Cognitive Unit 2.) but also that **A** did **a** particularly well (Cognitive Unit 5. of "praising"), better than the average, surprisingly, unexpectedly well: therefore, **S** is also surprised (Cognitive Unit 6.- as we shall see later, surprise occurs any time a new event disconfirms some expectation). Moreover in praising,

different from approving, **S** not only has a good evaluation of the action **a**, but of the person **A** (Cognitive Unit 8.) as somebody who is better than the average (and this causal link is expressed by Cognitive Unit 9.). In conclusion, while the performative of approving contains only the emotion of happiness for the other's action, praising contains not only happiness but also surprise. And these emotions can be detected in the face of approving and praising.

3 Emotion in the Expression of Performatives

That some affective states are contained in some performatives is true not only on the level of the cognitive structure of performatives but also on the level of how they are conveyed through facial expression. If we turn to the signal side of performative faces we can see that some Action Units that typically convey affective states take part, in fact, in the facial expression of performatives, and sometimes they even show a one-to-one correspondence with their affective cognitive units. For instance, in the imploring face the inner parts of the eyebrows are raised (see Figure 1 (a)); which is also the typical expression that characterises sadness. Similarly, the face in a peremptory order performs a frown (inner parts of the eyebrows lowered and closer), which is also a typical expression of anger (see Figure 1 (b)). A warning face contains an expression of worrying, with eyebrows closer and making wrinkles on the forehead (see Figure 2 (a)); a suggesting face contains an expression of uncertainty, with slightly raised eyebrows (see Figure 2 (b)), an approving face contains a slight smile of happiness with raised cheeks (see Figure 3 (a)), while a praising face contains the eyebrow raising and opened eyes typical of surprise (see Figure 3 (b)).



Fig. 1. (a) Imploring eye

(b) Ordering eye

4 Raising Eyebrows: Not only Surprise

So far we have shown that emotional information is generally contained in the expression of a performative, that is, the specific communicative goal of a sentence or other kind of communicative act. But emotional content is also present in other communicative material. In order to explore this topic, we do not start anymore from the meaning side but from the signal side of facial expression. Take

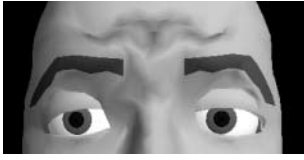


Fig. 2. (a) Worrying eye



(b) Suggesting eye



Fig. 3. (a) Approving eye



(b) Praising eye

a specific facial signal, the raising of eyebrows, and see what different meanings it may convey. This facial signal occurs in different situations and may bear the following different meanings:

1. When facing an unforecast situation it means “I am surprised”;
2. When listening to an interlocutor and wanting to show doubtfulness, it means incredulity, and it could be paraphrased as: “it’s very difficult to believe that!”
3. While uttering an adversative adverb or conjunction, like ‘but’, ‘however’, ‘instead’, ‘on the contrary’, it bears the same adversative meaning, that is, it warns the interlocutor not to draw the inference that would be most plausible from the preceding words. It equals warning: “it is not what you might think, the opposite is true in fact”;
4. In plain conversation or argumentation, it puts the emphasis on the word being uttered, thus meaning something like: “this is the most important thing in my sentence; this is what I really want you to understand”;

In a paper on the expressive movements of the eyebrows, Ekman [12] distinguishes these four seemingly very different situations: the first possibly holding even in absence of communicative interaction, the second in the case of the interlocutor in a conversational setting, the third and the fourth typical of the behavior of the Speaker in conversation. He argues that the upward and downward movements of the eyebrows are at work both in emotional signals - the expression of emotions of surprise, anger or fear - and conversational signals - interrogative expressions, emphasis and approving signals. His account seems to imply a strong and inexplicable polysemy in the same signal - for example, the same raising of the eyebrows may be on the one side an expression of surprise, on the other side a signal of emphasis.

On the basis of Ekman's account, these meanings do not seem to share a common core of meaning. We believe the opposite. That is, in our view, it is more economic and plausible to think that the different meanings (including the emotional meanings) of the same signal (surprise, here) have a part of their meaning in common.

Let us see now what might be the common core of meaning that the raised eyebrows share in the four cases above: 1. **surprise**, 2. **doubt** and **perplexity**, 3. **adversative word**, 4. **emphasis**.

First of all, let us see what is surprise and what is its function from a cognitive point of view.

Surprise is an emotion that occurs every time some expectation is disconfirmed, because some event has occurred that cannot be inferred by previous beliefs [8].

But why is the disconfirm of an expectation so relevant in human life as to give rise to an emotion, and why is it so important to communicate it?

To be aware that something is different from expected is important, from an adaptive point of view, particularly for the human animal, whose survival very heavily depends on its cognitive capacity. For humans, the most important resources to reach their goals are not as much their physical powers as their beliefs. This is why humans do not only have pragmatic or social goals, like the goal of performing actions well or the goal of being loved by other people; they also have epistemic goals, that is, the goal of acquiring as many beliefs as they can and of having one's beliefs reliable, integrated and linked with each other in cognitive networks [9]. Among epistemic goals humans also have the goal of generating inferences, of drawing expectations about events: this is essential not only to widen the range of their beliefs, but also to link beliefs with each other, in order to feel more sure of them; moreover inferences, and particularly expectations about future events, are necessary in planning the future course of action. But since, at the same time, inferences are by definition not completely reliable, and may be disconfirmed by the actual course of events, humans also have the epistemic goal of being particularly alerted any time some expectation is contradicted. Now, when some goal is particularly important from an adaptive point of view, its being reached or thwarted is often signalled and monitored by an emotion; and surprise is just the emotion that is felt as an expectation is suddenly disconfirmed. But again, since an unexpected belief jeopardises the whole structure of human knowledge, in that it cannot be inferred from previous beliefs, each new unexpected event also triggers the search for further beliefs that allow humans to infer it. This is why the emotion of surprise causes them to raise eyebrows and to open their eyes wide: opening eyes wide is a way to widen the visual field, and then to see a larger amount of things, to catch a larger amount of beliefs. A surprised person is biologically programmed to be ready to catch as much visual information as she can: thus, the raising of eyebrows and the opening of eyes become an expressive signal of surprise. But why should this meaning be shared by the other situations above? Let us take them one by one,

and see what do these meanings have in common with surprise. This is quite easy to see when case 1. (surprise) and case 2. are compared.

The semantic structure of surprise may be represented as follows:

1. **S** comes to believe **p**
 2. **S** believes **q**
 3. **S** intends to infer **p** from **q**
 4. **S** cannot infer **p** from **q**
 5. **S** feels surprised
 6. 4. causes 5.
- Cognitive Units of surprise

Surprise occurs when **S** comes to believe something (Cognitive Unit 1.) that **S** cannot infer (Cognitive Unit 4.) by one's previous knowledge **q** (Cognitive Unit 2.), which is something that usually **S** wants (Cognitive Unit 3.). This impossibility to infer new information from old information causes the emotion of surprise (Cognitive Unit 5.).

Let us now see the semantic structure of case 2., the back channel of doubt or perplexity¹:

1. **S** believes that **A** has the goal that **S** believes **p**
 2. **S** believes **q**
 3. **S** cannot infer **p** from **q**
 4. **S** feels doubtful about **p**
- Cognitive Units of back channel of doubt

In the back channel of doubt, **S** is the former Addressee who is now the Sender of a back-channel message, something like "I can't believe what you are saying". **A** (the former Sender of a message) has the goal that **S** believes what **A** is saying (Cognitive Unit 1.), but **S** already believes **q** (Cognitive Unit 2.), from which **p** cannot be inferred (Cognitive Unit 3.); then **S** raises the eyebrows in order to show she is doubtful about **p** (Cognitive Unit 4.).

Therefore, showing **doubt** or **perplexity** (case 2.) as a back-channel signal to someone who is telling us something incredible is not so different from showing surprise, because an incredible belief is one I cannot infer from my previous knowledge, then something that disconfirms my expectations. One could object in fact that in such a situation the raising of the eyebrows is not necessarily accompanied by the opening of the eyes. This could be accounted for with the fact

¹ In general, in the representation of a back channel, we use **S** to denote the former Addressee since in the case of a back channel she is the Sender of a communicative act which can be verbal (for instance he can say 'hum-hum', 'I see', 'I can't follow') or nonverbal (head nod, smile, raising eyebrow). When the back channel is performed nonverbally, it often overlaps temporally with the Speech Act of the current speaker (previously called sender **S**). But since we analyze the communicative act of back channel the sender **S** is the one who is performing the back channel while the interlocutor, even if he is currently speaking, is at the moment the Addressee of this communicative act.

that the original need of having a wider field of view in the case of surprise, that was present in our ancestors, is not so compelling here; in cases like this, where the unexpected event is not happening in front of us but reported through other people's discourse, only the signal "raising eyebrows" remains to characterise the message.

Let us now come to case 3.: what is common between expressing surprise and accompanying an **adversative word**? Adversative words (like 'but', 'though', 'yet', 'nonetheless'...) signal that an expected belief or a plausible inference is not true in fact. Take the sentence: "She is blond, but she has black eyes". Here, from the first part of the sentence, "She is blond", you could infer that "she has blue eyes", since this is what usually occurs; therefore, in order not to let you draw this inference, which in this specific case is incorrect, I not only tell you the second part of the sentence, "she has black eyes", which provides the correct information; I also warn you not to draw the plausible but wrong inference by saying 'but', and/or by raising my eyebrows. In fact, 'but', as well as all adversative words and nonverbal signals, has just the function of stopping plausible but incorrect inferences [9]; it could then be paraphrased as: "What I am going to say in the next part of the sentence/discourse cannot be inferred, and even, the contrary could be inferred, from the previous part of the sentence/discourse",

The meaning of an adversative word could then be represented as follows:

1. **S** has the goal that **A** believes **q**
2. **S** has the goal that **A** believes **p**
3. **S** has the goal that **A** believes that **A** cannot infer **p** from **q**
4. **S** has the goal that **A** believes not **p** from **q**

Cognitive Units of adversative

In raising the eyebrows with an adversative function, **S** wants **A** to believe both the first and the second part of the sentence/discourse (Cognitive Units 1. and 2.), but at the same time acknowledges that **p** cannot be inferred from **q** (Cognitive Unit 3.), and even the contrary (not **p**) is true (Cognitive Unit 4.).

Thus, both surprise (case 1.) and adversative words (case 3.) share the same element of a disconfirmed expectation: in fact, as well as surprise occurs when an unexpected belief is assumed, also an adversative signal is provided when I warn you that an unexpected belief has to be assumed, one which contrasts with previous knowledge.

Finally, let us see what semantic relationship can be found between raising eyebrows out of **surprise** (1.) and out of **emphasis** (4.). As I want to emphasise a word or a clause in my sentence, I show a more awakened attention myself in order to ask more attention from my interlocutor, because the part of sentence I am uttering is the comment, my new contribution of information, and then the part I consider most important and most worth of attention. Raising the eyebrows in this case then means something like: "pay more attention than usual, because this is the newest, least obvious, most relevant information in what I am saying". In fact, there is a link between surprise and new information because any new information is potentially something we cannot infer. This is

why the element of something that cannot be inferred from previous knowledge is, at least potentially, present also in the case of emphasis. Emphasis can be represented like this:

1. **S** has the goal that **A** believes **p**
 2. **S** believes that **A** cannot infer **p** from **q**
 3. **S** has the goal that **A** pays attention to **p**
- Cognitive Units of emphasis

S wants **A** to believe **p** (Cognitive Unit 1.), but **S** believes **A** cannot infer **p** by previous knowledge (Cognitive Unit 2.), and ask **A** to pay attention (Cognitive Unit 3.).

To sum up, then, the semantic element of a new information that cannot be inferred from previous knowledge seems to be common to all four cases: surprise, doubt, adversative, and emphasis.

5 The Evolution of Meanings

The idea that a common core of meaning is shared by different readings of the same signal implies a hypothesis on the diachronic (historic and philogenetic) evolution of meanings of nonverbal signals. Our hypothesis is that a signal is displayed, at its first occurrences, while having a primitive meaning **1.**. Suppose for instance that the eyebrow raising firstly means that for the speaker some event is different from expected. Later, when that signal is produced in different contexts, thanks to each current situation some specific inferences (new beliefs) may be drawn from that signal: for instance, in a situation **1.** what is different from expected is an eclipse, where the usually always bright sun is obscured by something; in a situation **2.**, what is different from expected is an incredible thing that my interlocutor is telling me. Now, suppose these inferences, these additional elements of information, are always the same in the same class of contexts, at the extent that they come to be recurrently attached to the primitive meaning **1.**: in this case the new inferences become crystallised around the primitive meaning, that is, they are not produced only episodically, but they start to systematically form part of a new meaning, one made up by the primitive meaning plus the crystallised inferences. For example, in the class of contexts “natural phenomena” the signal exhibited may start to bear two beliefs (disconfirmed expectation + in natural phenomena), while in the class of contexts “verbal interaction” the same signal may bear again two beliefs, but the first is the same as in the former class, while the second is a different one (disconfirmed expectation + in what other people say). In this case the recurrent inferences have formed a new meaning, where the new beliefs, that were previously inferred a-systematically according to the context, are now permanently attached to the primitive meaning, in such a way as to form a new bunch of beliefs - in fact, a new meaning -, that is richer than the primitive meaning in that it bears more beliefs that it did before, but still includes the primitive meaning. This is, in

our view, one of the ways a new meaning evolves from a primitive meaning. According to this hypothesis, at the end of this process, by analyzing the different meanings of the same signal we should find that they all share one and the same part of meaning - a common core of meaning - , but they differ from one another in that each different meaning has one or more parts that do not overlap with the other meanings. This structure of polysemic signals may be represented in this way:

MEANING 1	MEANING 2	MEANING 3	MEANING 4
surprise	perplexity	adversative	emphasis
disconfirmed expectation	disconfirmed expectation	disconfirmed expectation	disconfirmed expectation
from natural phenomenon	from incredible statement	following part of sentence not inferable from previous part	new information
		do not draw most plausible inference	I ask you to pay attention

6 **Meaning-Based versus Surface-Based Systems**

In building embodied agents with talking faces, agents capable of expressive and communicative behavior, we consider it important that the agent express his communicative intentions. Suppose an agent has the goal of communicating something to some particular interlocutor in a particular situation and context: he has to decide which words to utter, which intonation to use, and which facial expression to display. If the agent moves only his lips to talk but uses no other signals (no intonation to mark an accent or the end of an utterance, no facial expression, no change in the gaze direction, no hand gesture, and so on), the user might soon loose the impression of dialoguing with an embodied agent. Moreover, the user might have a hard time understanding what the agent is saying, since no communicative nor affective information other than plain text will be present in his discourse. Such an interaction will lack the necessary elements to provide a natural interaction. So the creation of an embodied agent requires the agent to be able to exhibit not only mere expression of emotion but also other expressions with subtle communicative functions. The theory presented in this paper based on the analysis of various communicative acts provides the necessary formalism for the construction of such an agent. It provides the foundations of a way to represent several complex elements (belief, goal, emotional state, power relationship, social relationship and so on) that are part of the mental state of the agent, as well as the inference rules that the agent uses to deduce his verbal and nonverbal behaviors. Through a fine-grain representation of mental states, we

can make agents with enough knowledge and flexibility as to adapt to different contexts and to exhibit different personalities and attitudes². A first approach of this work is being integrated in a discourse generation program [23] based on Mann, Matthiesen and Thompson's 'Rhetorical Structure Theory' (RST) [17]. The leaves of the RST tree can be either verbal or nonverbal signals. A goal-media prioritising module determines which media (face, gaze, speech) will be used, while the function of the synchronizer module is to compute the occurrence and duration of each signal (e.g. a raising eyebrow may coincide with individual words while emotional display may span several clauses). We are conscious we are at the premises of such a creation but we think that animating an agent from a system that is only based on surface signal (i.e. facial expression) and not on meaning (i.e. communicative act) will not be able to embed the naturalness and richness of human communication.

7 Emotional Expression as the Core of Communication: Conclusion and Further Research

In this paper we have shown that an emotional component of meaning (sadness, happiness, worry...) is present in communicative signals other from mere expression of emotion, namely the expression of the performative; on the other hand we have shown that a core of meaning which is contained in the expression of an emotion (surprise) is not present only in that emotion per se but also in other kinds of communicative signals (emphasis, back channel of doubt, adversative). The raising of the eyebrows, with its meaning of surprise, that is of a disconfirmed expectation, may not only be found in the expression of the emotion of surprise, but it may also accompany adversative conjunctions or emphasise the comment of a sentence; and all of these communicative signals imply something new, different from expected, and then surprising. We argue that a construction of animated faces based simply on the representation of surface signals can not catch the subtleness and richness of the possible meanings that underly communicative facial expressions. Instead a componential view of facial expression could be particularly apt to produce a modular and flexible expressive capacity of faces, making them able to exhibit emotional expression either by itself or as part of other interactional communicative signals.

References

1. G. Ball and J. Breese. Emotion and personality in a conversational agent. In S. Prevost J. Cassell, J. Sullivan and E. Churchill, editors, *Embodied Conversational Characters*. MITpress, Cambridge, MA, 2000. 183
2. J. Bates. Realism and believable agents. In *Lifelike Computer Characters'94*, 1994. 183

² For more details on the definition of context, and on how context, personality, power relationship are all variables used in determining the communicative behavior of the agent, the interested reader is referred to [21,22].

3. J. Beskow and S. McGlashan. Olga - A conversational agent with gestures. In *Proc. of IJCAI'97 - Workshop on Animated Interface Agents - Making them intelligent*, Nagoya, Japan, August 1997. Morgan-Kaufmann Publishers, San Francisco. 183
4. J. Cassell, J. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsón, and H. Yan. Embodiment in conversational interfaces: Rea. In *CHI'99*, pages 520–527, Pittsburgh, PA, 1999. 183
5. J. Cassell, C. Pelachaud, N. I. Badler, M. Steedman, B. Achorn, T. Becket, B. Derville, S. Prevost, and M. Stone. Animated conversation: Rule-based generation of facial expression, gesture and spoken intonation for multiple conversational agents. In *Computer Graphics Proceedings, Annual Conference Series*, pages 413–420. ACM SIGGRAPH, 1994. 183
6. J. Cassell and K. Thórisson. The power of a nod and a glance: Envelope vs. emotional feedback in animated conversational agents. *Applied Artificial Intelligence*, 13(3), 1999. 182
7. C. Castelfranchi, F. de Rosis, R. Falcone, and S. Pizzutillo. A testbed for investigating personality-based multiagent cooperation. In *European Summer School of Logic, Language and Information*, Aix-en-Provence, France, 1997. 183
8. C. Castelfranchi and I. Poggi. 'beh!' e 'beh...'. Analisi semantica di una interiezione italiana. In I. Poggi, *Le interiezioni. Studio del linguaggio e analisi della mente*. Boringhieri, Torino, 1981. 188
9. C. Castelfranchi and I. Poggi. Bugie finzioni sotterfugi. In *Per una scienza dell'inganno*. Carocci, Roma, 1998. 188, 190
10. R. Conte and C. Castelfranchi. *Cognitive and Social Action*. University College, London, 1995. 183
11. P. Doyle. When is a communicative agent a good idea. In *Workshop on "Communicative Agents: The Use of Natural Language in Embodied Systems"*, Third International Conference on Autonomous Agents, Seattle, May 1999. 182
12. P. Ekman. About brows: Emotional and conversational signals. In M. von Cranach, K. Foppa, W. Lepenies, and D. Ploog, editors, *Human ethology: Claims and limits of a new discipline: contributions to the Colloquium*, pages 169–248. Cambridge University Press, Cambridge, England; New-York, 1979. 187
13. P. Ekman and W. Friesen. *Unmasking the Face: A guide to recognizing emotions from facial clues*. Prentice-Hall, Inc., 1975. 183
14. P. Ekman and W. Friesen. *Facial Action Coding System*. Consulting Psychologists Press, Inc., Palo Alto, CA, 1978. 184
15. T. Koda and P. Maes. Agents with faces: The effects of personification of agents. In *HCI'96*, August 1996. 182
16. Y. Lee, D. Terzopoulos, and K. Waters. Realistic modeling for facial animation. In *Computer Graphics Proceedings, Annual Conference Series*, pages 55–62. ACM SIGGRAPH, 1995. 183
17. W. C. Mann, C. M. I. M. Matthiessen, and S. Thompson. Rhetorical structure theory and text analysis. Technical Report 89-242, ISI Research, 1989. 193
18. C. Pelachaud, N. I. Badler, and M. Steedman. Generating facial expressions for speech. *Cognitive Science*, 20(1):1–46, January-March 1996. 183
19. S. M. Platt. *A Structural Model of the Human Face*. PhD thesis, University of Pennsylvania, Dept. of Computer and Information Science, Philadelphia, PA, 1985. 183, 184
20. I. Poggi and E. Magno Caldognetto. *Mani che parlano. Gesti e Psicologia della comunicazione*. Padova: Unipress, 1997. 183
21. I. Poggi and C. Pelachaud. Performative faces. *Speech Communication*, 26:5–21, 1998. 183, 184, 193

22. I. Poggi and C. Pelachaud. Facial performative in a conversational system. In S. Prevost J. Cassell, J. Sullivan and E. Churchill, editors, *Embodied Conversational Characters*. MITpress, Cambridge, MA, 2000. 183, 193
23. I. Poggi, C. Pelachaud, and F. de Rosis. Eye communication in a conversational 3d synthetic agent. *Special Issue on Behavior Planning for Life-Like Characters and Avatars of AI Communications*, 2000. 183, 193
24. B. Reeves and C. Nass. *The media equation: How people treat computers, television and new media like real people and places*. CSLI Publications, Stanford, CA, 1996. 182
25. A. Takeuchi and T. Naito. Situated facial displays: Towards social interaction. In *Proceedings of ACM CHI'95 - Conference on Human Factors in Computing Systems*, volume 1, pages 450–455, 1995. 182
26. K. R. Thórisson. Layered modular action control for communicative humanoids. In *Computer Animation'97*, Geneva, Switzerland, 1997. IEEE Computer Society Press. 183
27. J. H. Walker, L. Sproull, and R. Subramani. Using a human face in an interface. In *Human Factors in Computing Systems*, pages 85–91, April 1994. 182

Relating Personality and Behavior: Posture and Gestures

Gene Ball and Jack Breese

Microsoft Research
geneb@microsoft.com

Abstract. We have constructed a Bayesian model relating personality and emotion to externally observable behaviors, designed to be useful in generating natural and convincing communicative behaviors in a conversational agent. The same model can be used to diagnose the internal emotional state and personality type of the human user. This paper briefly recounts the motivation and structure of the overall model, and then considers the relationship between personality type and posture and gestures in more detail. As is well established in the psychology literature, people recognize characteristic body motions as reliable indicators of the personality type of others. We review the most likely sources of evidence for those judgments, and consider the feasibility of generating behavior in an animated computer agent that presents a consistent personality, as well as the more difficult task of recognizing personality type based on body movement.

1 Introduction

We are moving toward the day when we will regularly converse with our computers in natural spoken conversation. Creating an animated visual character will be a common way to lend substance to our conversational computers. These characters will inevitably become a significant social presence, and for many people, will eventually become long-term companions with which (whom?) they share much of their day-to-day activity.

The considerable challenges of creating conversational computers can be characterized in three stages:

- Usefulness: making conversational systems which can competently provide some desirable service;
- Usability: ensuring that their communication skills are sufficient to allow robust and efficient interactions; and
- Comfort: creating conversational partners that can fulfill our deeply ingrained expectations about how human conversations take place.

In many fundamental ways, people respond psychologically to interactive computers as if they were human. Nass and Reeves [14] have demonstrated that strong social responses are evoked even if the computer isn't presented as an explicitly anthropomorphic embodied agent. They suggest that humans have evolved to accord special significance to the movement and language produced by other people, and cannot avoid responding to the communication of 20th century technology as if it were coming from another person.

In order to build *comfortable* systems, we will need to understand the psychological reality and significance of these effects and to adapt our computer systems to the needs of our users. Whenever the behavior of a conversational system is not completely constrained, the particular choices that it makes among available behaviors will be interpreted by the user as expressions of its "personality" [10]. Even if we attempt to discourage this interpretation by eliminating all extraneous behavior, we will simply project a different personality: that of a reserved, businesslike, and unfriendly assistant. Therefore, explicit attention to the social aspects of computer interaction will be necessary in order to avoid degrading the user's experience by generating unnatural or disconcerting behaviors.

Personality traits that might be appropriate for an embodied agent are illustrated by the following plausible (but imaginary) responses from the agent to a user:

- This doesn't seem to be going so well... shall we try again? (friendly)
- Excuse me... You asked for a reminder if you were on the Web for more than 20 minutes... (submissive)
- You'd better get back to work now. (arrogant)
- Gee, that was pretty frustrating, wasn't it? (unassuming)
- Take a break now, if you want your wrists to get better! (authoritarian)

In these examples, the agent's linguistic expression is the clearest indicator personality, but if the character is to seem natural and believable, it needs to be accompanied by appropriate non-linguistic signals as well.

Since personality traits are (by definition) aspects of behavior that change only very slowly, if at all, consistency is particularly important; otherwise the user will perceive the character as abnormal and will tend to find the interaction unpleasant. There is also some evidence [12] that the user's experience can be positively affected by the proper selection of an agent's personality type.

In order to project a recognizable and consistent personality type, conversational systems must be able to synthesize behaviors that are appropriate for that type. In addition, it may be useful to be able to diagnose the user's own personality type (by observing their characteristic behaviors), in order to adjust the agent's personality and produce a more comfortable working partnership [14].

1.1 Related Work

While many animated systems have been authored to convey a distinct personality through pre-defined animation or audio effects, few have attempted to explicitly select a personality type and then generate appropriate behavior dynamically.

Perlin and Goldberg's Improv animation system [13] creates layers of small animations under the control of scripts and state variables. They have described the use of state variables to control mood and personality in their characters.

The Petz products by PF.Magic [15] are some of the most psychologically expressive systems. Their animated characters have distinct personality profiles and maintain an internal model of emotion that pervasively affects the choice of behaviors and their associated facial expression, posture, and vocalizations. The simulated animals are sufficiently psychologically believable (with only non-linguistic interaction) that their owners regularly develop strong emotional bonds with them.

Nass et. al. [12] demonstrated that the personality attribute of dominance can be efficiently communicated in a textual interaction by using strong assertions and commands, displaying a high confidence level, and always initiating the conversation. The same study showed that when subjects interacted with a computer that was similar in personality to themselves, they rated it as friendlier, more competent, and felt the interaction was more satisfying.

2 Modeling Personality

2.1 Personality

Psychologists recognize that individuals have long term traits that guide their attitudes and responses to events. They use the term personality to describe permanent (or very slowly changing) patterns of thought, emotion, and behavior associated with an individual. McCrae and Costa [9] analyzed Wiggins' five basic dimensions of personality, which form the basis of commonly used personality tests. They found that this interpersonal circumplex can be usefully characterized within a two dimensional space.

Therefore our system adopts a simple representation of personality, structured along the dimensions of:

- Dominance, indicating an individual's relative disposition toward controlling (or being controlled by) others, and
- Friendliness, measuring the tendency to be warm and sympathetic.

In our model, these two continuous dimensions are further simplified by encoding them as a small number of discrete values. Dominance is encoded in our model as Dominant, Neutral, or Submissive; and Friendliness is represented as Friendly, Neutral, or Unfriendly. Figure 1 shows the Dominance-Friendliness personality space and labels a few different personality types within it.

Our system also models emotional state, the short-term (often lasting only a few seconds) variations in internal mental state. As with personality, we focus on just two basic dimensions of emotional response [8] that can usefully characterize nearly any experience:

- Valence represents the positive or negative dimension of feeling, and
- Arousal represents the degree of intensity of the emotional response.

In our model, these two continuous dimensions are further simplified by encoding them as a small number of discrete values. Valence is considered to be Negative, Neutral, or Positive; similarly, Arousal is judged to be Excited, Neutral, or Calm.

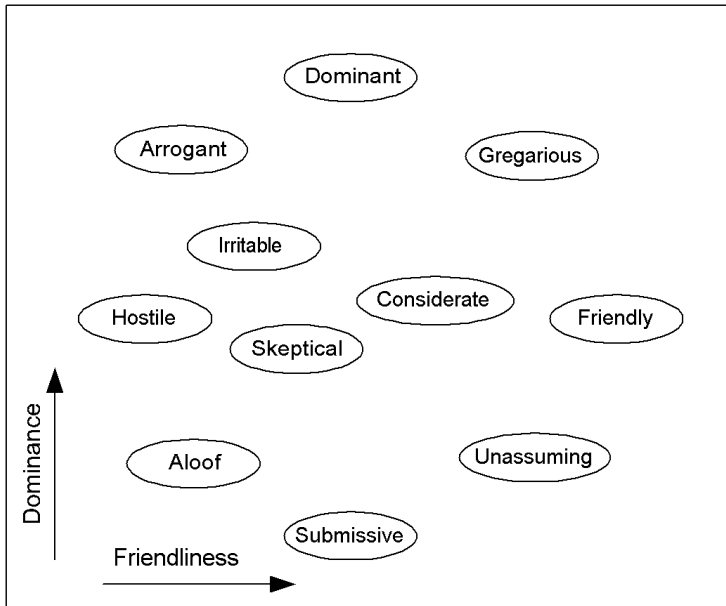


Fig. 1. Some named personality types within the dominance-friendliness space

2.2 Relating Internal State to Behavior

Given this quite simple, but highly descriptive, model of an individual's personality type and internal emotional state, we wish to relate it to behaviors that help to communicate that state to others. The behaviors to be considered can include any observable variable that could potentially have a causal relationship with these internal states. Personality type can be reliably measured by questionnaires such as the Myers-Briggs Type Indicator [11]. However, in normal human interaction, we rely primarily on visual and auditory observation to judge the personality type of others.

A computer-based agent might be able to use direct sensors of physiological changes, but if those measures require the attachment of unusual devices, they would be likely to have an adverse effect on the user's perception of a natural interaction. For that reason, we have been most interested in observing behavior unobtrusively, either through audio and video channels, or possibly by using information (especially timing) that is available from traditional input devices like keyboards and mice, but might be a good indicator of the user's internal state.

2.3 Bayesian Networks

We use a Bayesian network to model the relationships between personality and emotion and their behavioral expression. Bayesian networks [7] are a formalism for representing networks of probabilistic causal interactions that have been effectively applied to medical diagnosis [6], troubleshooting tasks [5], and many other domains.

Bayesian networks have a number of properties that make them an especially attractive mechanism for modeling emotion and personality:

- They deal explicitly with uncertainty at every stage, which is a necessity for modeling anything as inherently non-deterministic as the connections between emotion/personality and behavior.
- The links in a Bayesian network are intuitively meaningful, since they directly represent the connections between causes and their effects. The conditional probabilities of the possible outcomes must be estimated, but the trends of the effects are often readily apparent.

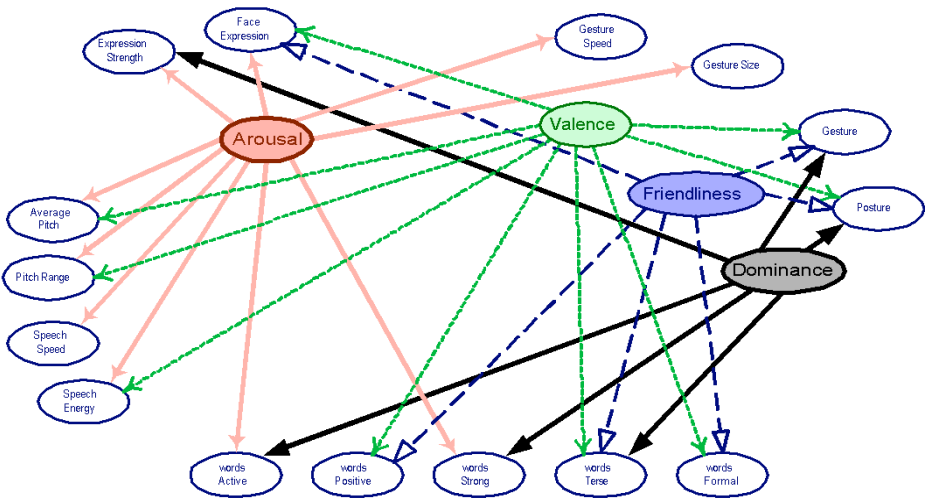


Fig. 2: Bayesian network representing model of personality and emotion

- The network can be extended easily by introducing a new behavior node and linking it to the internal states that most strongly affect it. The probability distributions of the new links are independent of the rest of the model.
- They can be used both to calculate the likely consequences of changes to their causal nodes, and also to diagnose the likely causes of a collection of observed values at the dependent nodes. This means that a single network (and all of its parameters) can be used both for simulating believable behavior in an agent, and for recognizing the significance of the user's behavior.

The dependency graph of our model of personality and emotion is shown in Figure 2. The behaviors currently represented include facial expressions, linguistic style, vocal expression, and body movements. We focus here on the connections between personality type and body posture and gestures; the remainder of the model is discussed in more detail in Ball [1].

3 Gesture and Posture

Our observations of body movements tell us a great deal about the personality and emotional state of others [3].

Dominant personality traits are strongly communicated by postures and gestures that demonstrate a readiness for aggressive action. Body positioning that emphasizes personal size, a strong upright (“confident”) posture, hands placed on the hips, and directly facing the listener all convey dominance. Bending back and tilting the head back suggests arrogance and disdain [4]. A relaxed, asymmetrical positioning of the body conveys fearlessness, which also suggests dominance. Gestures like reaching forward with palms down, slapping a surface or focusing a direct unwavering gaze at another, also communicate a dominant personality type.

By contrast, submissive personalities tend to adopt postures that minimize size, such as slouching or kneeling, and position their bodies at an angle. Submissive gestures include bowing (showing harmlessness), gazing down, tilting the head to the side, reaching out with palms up, and shrugging the shoulders.

A friendly personality (as well as positive emotional valence) is communicated by postures and gestures that increase accessibility to a conversational partner. These include leaning forward, directly orienting the body, placing arms in an open position, and a direct gaze (when coupled with a forward lean and smile). In addition, submissive displays, such as shoulder shrugs and tilted head, indicate harmlessness and signal friendly intent.

Although posture and gestures effectively communicate personality type, that is not their sole communicative function. Human conversations include a variety of face, hand, and body movements that play specific and significant roles in communication, and in the maintenance and synchronization of the communication channel itself. A conversational agent cannot produce a natural and comfortable interaction without properly generating these signals [2]. The emotion and personality dimensions of body movement are best viewed as background behaviors behind the conversational gestures that carry more specific meaning.

Therefore, our model estimates the relative likelihood of a variety of specific postures and gestures, depending on the personality type (and emotional state) of the agent. When using the model to control the behavior of an animated agent, the system sets the values of the Dominance and Friendliness nodes (along with the emotional states) and evaluates the Bayesian network. This results in a probability distribution indicating which postures and gestures are most likely to convey the desired personality. These distributions can then be sampled to select a sequence of background postures and gestures. These actions must be integrated with the stream of emblematic (“thumbs up”), propositional (“it was this big”), iconic (depicting some

feature of a referent), metaphoric (abstract symbols), deictic (pointing), and beat (synchronization) gestures that accompany normal conversational speech [2].

The timing of background body movements must also be determined: once a particular gesture has been chosen (e.g. hands on hips), the duration of that behavior is determined, and then aligned with any foreground gestures that may be accompanying speech. Gesture speed (and frequency) are currently based on emotional arousal levels only.

4 Conclusion

We have constructed a Bayesian network model that estimates the likelihood of specific body postures and gestures for individuals with different personality types. By sampling the resulting probability distributions, we can generate a sequence of body movements that seem to be consistent with the selected personality type. Although these motions convey a distinct impression of personality, they are not convincingly natural, because they are not (as yet) coordinated with appropriate speech-synchronized communicative gestures.

Decoding user body movements from visual input is a much more difficult problem. It will first become feasible to detect indications of arousal based on the size, frequency, and speed of body movements. More detailed recognition will first require the reliable identification of specific visual gestures, already a difficult problem. Then it will be necessary to determine which gestures are tightly coupled with speech and which are more appropriately interpreted as indications of personality. This discrimination will require at least some level of understanding of the user's speech.

Therefore, in the near term, analysis of user personality will probably rely primarily on linguistic analysis and such factors as typing speed. While these methods are likely to be somewhat unreliable, Nass and Reeves point out that users seem to appreciate a system that changes its communication style in an attempt to accommodate their personality, and that a mistaken adjustment does *not* cause them to strongly dislike the system [14]. Thus a conversational agent that makes its best estimate of the user's personality type and then adopts a similar style, is likely to be preferred over one that doesn't adapt at all.

References

1. Ball, G. and Breese, J. (forthcoming). "Emotion and Personality in a Conversational Agent." In J. Cassell (Ed.), *Embodied Conversational Agents*. Cambridge, MA: MIT Press.
2. Cassell, J. (forthcoming). "Embodied Conversation: Integrating Face and Gesture into Automatic Spoken Dialogue Systems." In Luperfoy (ed.), *Spoken Dialogue Systems*. Cambridge, MA: MIT Press.
3. Collier, G. 1985. *Emotional Expression*. Hillsdale, NJ: Lawrence Erlbaum Associates.

4. Givens, David. (forthcoming). *The Nonverbal Dictionary Of Gestures, Signs and Body Language Cues*. Center for Nonverbal Studies. Preview available at: <http://members.aol.com/nonverbal2/diction1.htm>
5. Heckerman, D., Breese, J., and Rommelse, K. 1995. "Troubleshooting under uncertainty." *Communications of the ACM*. 38(3): 49-57.
6. Horvitz, E. and Shwe, M. (1995). "Melding Bayesian inference, speech recognition, and user models for effective handsfree decision support." In *Proceedings of the Symposium on Computer Applications in Medical Care*. IEEE Computer Society Press.
7. Jensen, F. V. 1996. *An Introduction to Bayesian Networks*. New York: Springer-Verlag.
8. Lang, P. 1995. "The emotion probe: Studies of motivation and attention." *American Psychologist*. 50(5): 372-385.
9. McCrae, R. and Costa, P. T. 1989. "The structure of interpersonal traits: Wiggin's circumplex and the five factor model." *Journal of Personality and Social Psychology*. 56(5): 586-595.
10. Moffat, D. 1997. "Personality Parameters and Programs." In Trappl and Petta (Eds.), *Creating Personalities for Synthetic Actors*. (pp. 120-165). Berlin: Springer.
11. Myers, I. B. and McCaulley, M. H. 1985. *Manual: A Guide to the development and use of the Myers-Briggs Type Indicator*. Palo Alto, California: Consulting Psychologists Press.
12. Nass, C., Moon, Y., Reeves, B., and Dryer, C. 1995. "Can Computer Personalities Be Human Personalities?" *Journal of Human-Computer Studies*. 43:223-239.
13. Perlin, K. and Goldberg, A. 1996. "Improv: a system for scripting interactive actors in virtual worlds." In *SIGGRAPH '96. Proceedings of the 23rd annual conference on Computer graphics*. (pp. 205-216). ACM.
14. Reeves, B. and Nass, C. 1995. *The Media Equation*. New York: CSLI Publications and
15. Stern, A., Frank, A., and Resner, B. 1998. "Virtual Petz: A Hybrid Approach to Creating Autonomous, Lifelike Dogz and Catz." In *Proceedings of the Second International Conference on Autonomous Agents, AGENTS98*. (pp. 334-335). ACM Press.

Affective Natural Language Generation

Fiorella de Rosi¹ and Floriana Grasso²

¹ Dipartimento di Informatica - Università di Bari, Italy

`derosis@di.uniba.it`

² Department of Computer Science - University of Liverpool, UK

`floriana@csc.liv.ac.uk`

1 Introduction

The automatic generation of natural language messages (or NLG) has been employed in many computer systems and for various purposes, which go from providing information on a particular subject, to instructing on how to perform some -complex- action, to arguing about a particular claim (for an overview, see [24]).

In the majority of systems that have been designed in the last decade, these messages are adapted, in their content and presentation style, to the context in which they have to be applied: that is, to the Hearer's characteristics (represented in a "mental model" of the Hearer), to the application domain and, generally in a more implicit way, to the Speaker's characteristics. Adaptation is based on "strong" assumptions about the Hearer's mental state and the way this state is influenced by communication of each individual component of the message, and by understanding of the relationships among these components. The aspects of the mental state that are represented, in the large majority of cases, are the Hearers' beliefs and knowledge of domain topics and their goals about the domain state; in some cases, this may extend to representing other aspects, such as the ability to perform domain actions, the interest towards topics, the preference towards domain states. When a Speaker's model is represented as well, this includes second-order beliefs and goals of the same type.

Generally seen as informative tools, these systems give little space to representation of less rational aspects, such as emotions, of both Speaker and Hearer. However, natural language communication is influenced by a number of factors, of which the more rational ones constitute only a subset. Especially when communication occurs in "delicate" scenarios, such as a medical setting, the Hearer cannot be expected to coolly react to what is being said. Many studies in the health behaviour research have shown that patients' attention and understanding is highly affected by their emotional involvement [1,26]. It therefore appears worthwhile to investigate whether, when and how emotions, personalities and other extra-rational factors should be taken into account when designing a NLG tool.

By paraphrasing Rose Picard [22], we define *Affective Natural Language Generation* as "NLG that relates to, arises from or deliberately influences emotions or other non-strictly rational aspects of the Hearer": these aspects (denoted with the generic term of "attitudes") include personality traits, emotions and highly-placed values. We argue for the need for affective NLG by considering

an example from medical explanation texts, in which affect plays a particularly relevant role. We claim that Affective NLG requires a revision of the generation methods and of the knowledge sources (about the domain and about the Hearer) that are employed in this process, and, from analysis of the presented example and of the literature, we draw some conclusions about the way NLG methods might be revised to produce more affective texts.

2 A Case Study

We consider two examples from a corpus of explanations about drug prescription [2]. The corpus was collected by presenting a set of scenarios to several doctors, who were then asked to make a drug prescription and to support this prescription with an explanation addressed to the patient, to a nurse and to a colleague.

The two examples presented here are both addressed to the patient, as these were the versions of the explanation conveying more affect, and refer to the following scenario:

“A 44 years-old alcoholic man lives on a poor diet in a cold and damp lodging house. He now complains of a persistent cough, occasionally bringing up blood. His appetite is poor, and he thinks he is losing weight. Examination and chest X-rays, followed by other tests, confirm the diagnosis of tuberculosis”.

The explanation texts provided by the two doctors are shown in Fig. 1. Doctor D1 (column 2) provides a shorter text, Doctor D2 (column 3) a longer one. The two texts do not really differ in the overall structure: they both start with illustrating the disease, to then anticipate that treatment will be long and cannot be avoided; after synthesizing the main aspects of treatment (number of drugs and intake modalities), the texts end with mentioning the main accompanying measures that the doctor is going to undertake.

Both texts employ some typical *affect-conveying* techniques (italicized text in the figure). However, it is the quantitative and qualitative difference in the way these techniques are employed which makes the two texts really different:

- D1’s text is mainly focused on convincing the patient to *follow the treatment*, by relying on his sense of responsibility: treatment is long and necessary (4) and has to be done every day, for at least three months (7).
- D2’s text combines the goal of *motivating the patient* with the goal of *re-assuring him* about the perspectives of success of treatment. This text is longer, first of all because it includes many more justifications:
 1. of diagnosis, and the way it is linked to available evidence (1),
 2. of the reason why a combination of drugs is needed (6),
 3. of the reason why a notification of the disease is required (not from the legal but from the public health viewpoint! 10), and
 4. of the circumstance that may justify a future hospitalization (11).

At the same time, the patient is constantly reassured about the efficacy of the treatment prescribed (2, 4, 5, 6) and about the health service participation to monitor his health status (11).

1	Mr Smith, <i>unfortunately</i> you have an infection of the chest which is called tuberculosis;	Well Mr Smith, we've got the tests back now and I think we might have mentioned your losing weight and being a bit poorly, <i>but we were a little bit worried</i> that you did have TB, or tuberculosis infection. And the tests have really confirmed that worry, it looks as if, on your chest X-ray, you've got active TB, and your sputum is also positive.
2		So, <i>the good news</i> is that we do have tablets that are <i>very effective</i> for treating TB,
3		<i>you do have to take several tablets</i> a day,
4	the problem, with this infection, is it <i>takes a very long time</i> to eradicate it from the body and therefore we have to undertake <i>quite a long course of treatment</i> which <i>it is essential</i> for you to fulfill for the full course.	and you will have to take them for some months <i>to get really over</i> this problem,
5		<i>but it is something we can do something about.</i>
6	Therefore, initially we are going to start off with 3, and <i>we may even</i> add 4 drugs when we talk to the microbiologist.	What we are going to do today is start you on two types of tablets, one a special antibiotic to <i>kill off</i> the TB bugs, the other ones are vitamins, because you can run a <i>bit short</i> of vitamins when you're on these tablets.
7	These drugs will have to be taken every day, and then we'll reassess the situation in three months time.	
8	I notice that you live on your own at the moment and	Now the other thing we've got to do is to look into why you've got this, <i>we're a little bit worried</i> that you're <i>maybe a bit undernourished</i> , and haven't been looking after yourself, and
9		we're going to see if we can speak to our social workers to see if they can improve on your surroundings and benefits.
10		What we're also going to have to do is notify you as a case of TB, because this is an infection and you might have caught it from one of your friends, or alternatively you could maybe cough on somebody and infect them.
11	I just wondered what you felt of coming into hospital over the next few days.	So, these are the main things we're going to be doing, and <i>we'll be keeping a close eye on you here at the clinic</i> to make sure you are getting better, and <i>we may need to bring you into hospital for a bit</i> if you don't progress as well as <i>we'd like</i> .

Fig. 1. Two explanation texts about drug prescription

The first effect of the difference in affect, in the two texts, is therefore *a difference of content*: more affect implies, in the considered example, more details¹. The second effect is a *difference of form*: in D2's text, many more affect-conveying terms are used:

- verbs: *kill off* the TB bugs, in segment 6; *a close eye on you*, in segment 11;
- adjectives: the *good news*, in segment 2;
- adverbs: a lot of *little bit* and some *very, really, maybe*, here and there, when needed to enhance positive aspects and mitigate negative ones.

¹ In other cases, more affect implies, on the contrary, *less* details, caused by the decision to omit, from the explanation, aspects that are difficult for the Hearer to accept or that might cause distress to the Hearer, as we will see later on.

Both texts (and especially D2's) are rather *redundant*: some topics are repeated with identical or equivalent wording. For instance:

- “it takes a very long time to eradicate”, and “we have to undertake quite a long course of treatment” (D1's text, segment 4);
- “The good news is that we do have tablets that are very effective for treating TB”, and “but it is something we can do something about” (D2's text, segments 2 and 5).

Other topics, on the contrary, are only touched on, without going into details: we will see some examples of this when we will talk about side effects, in particular.

Finally, both doctors employ the “first plural person” form of verbs, to give the patient the feeling that caring for the diagnosed disease is not something that concerns the patient alone, but it is something the doctor and the patient are managing together (in the case of D2, it is a “battle” they will fight jointly).

To conclude: affect manifests itself, in the examined example, both at the content and at the form level. Redundancy, inclusion of motivating and reassuring details, and elusion of demotivating topics are examples of the first level; use of enhancing or mitigating terms and of the first plural person form are examples of the second one. In other texts in the same corpus, we found several examples of a wise use of redundancy vs conciseness, to stress favourable or de-emphasize unfavourable information. We also found examples of summaries of topics that are relevant for correct treatment, that were introduced in the texts with the aim of reinforcing the patient's memory. Finally, we found a variation in the order of presentation of information items (for instance, side effects and drug administration) that was probably linked to their presumed impact on the patient's emotional state. In other examples of medical texts, other forms for expressing affect were employed: for instance, in a corpus of dialogues between doctor and patient in two different settings (a family planning service in Italy and an adolescent diabetic care service in the UK), several forms of *omission* were employed like elusion or deception [3].

All the examples we looked at, anyway, show that affect plays a crucial role in explanatory texts. We shall reflect, in the remaining of this paper, on how such a trait may be rendered in automated NLG.

3 Affect in Natural Language Generation

Apart from works which explicitly deal with (especially informal) argumentation [25,27,15,10], reflection on how texts may be generated, that give appropriate weight to affective components of both the Speaker's and the Hearer's mental state, has been rather episodic in the NLG community, and mainly with a focus on the wording of the text.

Among others, [12] focuses on the idea of “producing utterances empathetic to both the Speaker and the Hearer”, by offsetting “unpleasant” information and stressing “favourable” one, through detensifier and intensifier adverbs: the Hearer's mental model is enriched, to this purpose, with domain-related personal preferences, concerns, worries and related features. In [30] it is discussed how variables like the “social distance” between the Speaker and the Hearer, the

“power” that the Hearer has over the Speaker and the two agents’ “basic desires” might affect strategies for realizing a particular speech act. Similar considerations are made in other projects whose purpose is to build “life-like characters with personality and emotions”, as, for instance, in [18].

Very few works have dealt with the problem of organizing the text content when affect has to be taken into account. In his book on *Generating natural language under pragmatic constraints*, Hovy dedicates a Chapter to this subject by discussing, in particular, how to deal with situations in which there is a “potential conflict” or an “agreement” between the affective implications of the text and the Hearer’s opinion [13, Chapter 4]. He suggests to apply content-related and form-related techniques, the first ones aimed essentially at avoiding “sensitive issues” or at de-emphasizing them through evasion, selectivity or other means, the second ones aimed at enhancing positive aspects and mitigating negative ones, through appropriate word choices. Hovy’s proposal to intervene on the text content through some form of evasion, elusion or “deception by masking” will remain an exception for some time. Kölln shifts this research field’s focus towards the planning step, by suggesting to consider some Hearer’s attitudes such as “subjective preferences” towards relevant concepts in the domain under consideration, in deciding the text content in concept explanation generation [16].

Some other researchers have been concerned with how emotional factors affect the structure of the sentence. One of the most notable contributions is Marilyn Walker’s study on the role of repetitions in texts [28,29]: this study goes in the opposite direction of the prevalent tendency in sentence planning, whose main focus is on *aggregation* [14]. In a recent survey, Reape and Mellish [23] report several definitions of this NLG task found in the literature, for instance: “redundancy elimination, abbreviation, text structure combination for concise and coherent text generation” etc. These goals may be achieved by language-dependent or language-independent techniques and may focus on conceptual, semantic or rhetorical aspects of the text. There are no considerations, in this paper like in others in the same field, on when aggregation should *not* be made. By illustrating the benefits of repetitions in a text, Marilyn Walker, on the contrary, indirectly indicates some circumstances in which redundancy is a value rather than a vice and therefore delimits the role of aggregation. We came to similar conclusions in our analysis of medical explanation texts and dialogues, in [6]: as the emotional state of the Hearers influences their reaction to the communicated message, assumptions that generally justify the avoidance of repetitions are likely to be violated in affective texts, and NLG methods should be revised accordingly.

Apart from these and few other examples, we lack, to our knowledge, a comprehensive research which fully accounts for how the goal of producing affective texts influences the various stages of the NLG process. In the next section we try to make a step in this direction, by sketching some ideas based on the literature and on our experience.

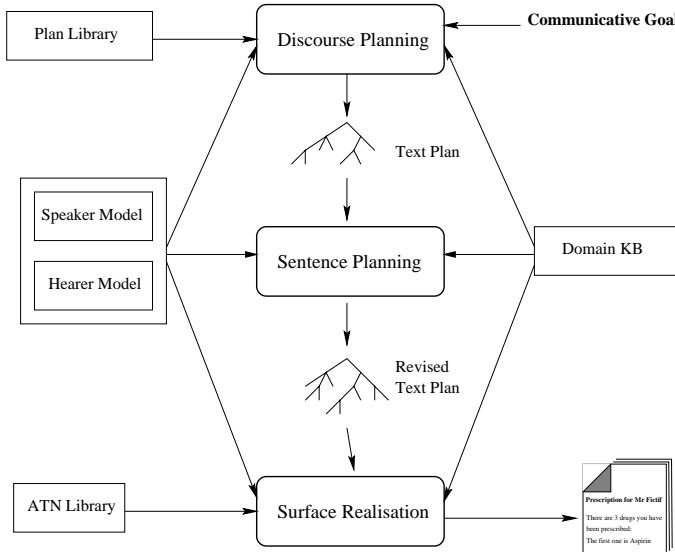


Fig. 2. Architecture of the NLG system

4 Considerations on How to Generate Affective Texts

If affect influences the information content and order and the way that communicative acts are realized, then all phases of the text generation process should be influenced by consideration of emotional factors. We assume a classic pipeline architecture for NLG systems, consisting of the phases of discourse planning, sentence planning and surface realization [24].

Fig. 2 illustrates this architecture as developed in our system: the Plan Library and the Augmented Transition Networks (ATN) library are knowledge sources employed in discourse planning and surface realization. A Speaker and a Hearer Model enable a double adaptation of the message produced; a Domain Knowledge Base (DKB) provides information on concepts, objects and other topics in the domain to which the explanation refers.

4.1 Discourse Planning

Discourse Planning is “the process of imposing ordering and structure over the set of messages to be conveyed” [24]. The result of this process, the text plan, is usually represented as a tree, whose leaves are the information content of the text and whose internal nodes have some representation of how parts of the text relate to one another.

Information content and order of the discourse items are often established by applying some planning algorithm to a Library of *plan operators*. The structure of the operators that have been applied in various discourse planning experiences varies slightly ([21] contains probably the most complete and clear introduction to this subject). We will refer, in this paper, to the structure of the operators that

we employed in our tool for generating explanations about drug prescriptions [7]. In these operators:

- the *Header* specifies a discourse segment purpose;
- the *Domain Constraints* specify conditions on the Domain Knowledge base that must be satisfied for the operator to be applied;
- the *Intentional Constraints* specify conditions on the Hearer mental state that should be satisfied for the operator to be applied. As opposed to domain constraints, intentional constraints can in turn be planned for, if not satisfied, through a pattern-matching with the Effects of other operators;
- the *Effects* specify the changes produced in the mental state of the Hearer by the communicative action(s) activated by the operator;
- the *Decomposition* specifies how the communicative goal mentioned in the Header may be decomposed into sub-goals, and the role that each of these subgoals plays in the Rhetorical Relation (RR) attached to the operator²;
- the *Rhetorical Relation* specifies the above mentioned RR name.

Figure 3 provides an example of one of these operators, that aims at describing a side effect (?x) of a drug (?y). The application constraint of this operator is that the Hearer (H) does not already know these side effects (NOT (KnowAbout H ?x)). Its effect is that the Hearer will come to know them. The goal specified in the Header may be decomposed into four subgoals (in this case, primary communicative actions) which are linked by a RR of *Elaboration Object-Attribute*; the goal to inform about the side effect name (Inform S H (Sign ?x)) is the nucleus of this RR; the other items in the Decomposition are satellites.

In our generator of explanations about drug prescriptions, about 100 operators were employed to plan the message structure. Although this method enabled us to achieve a double adaptation of generated texts (to the Speaker and to the Hearer), these texts did not show the same level of empathy that we had found in the doctors' corpus of explanations.

To produce more affective texts, such a method should be revised throughout. First of all, the strategy to employ in achieving the communicative goal that is formalised in a plan operator has to consider not only the Hearer's experience or knowledge but also other, less rational factors, such as their desires, interests, fears, satisfaction and so on. Moreover, these factors should be 'graded' rather than being considered as dichotomic variables. This requires revising the Intentional Constraints and the Effects slots of plan operators.

Intentional Constraints. As said before, this is the slot in which the decision of when to include a topic in the message is formalised. In traditional text planners, this decision is driven by the need to avoid redundancy in the text, by avoiding to talk about what the user already knows. This originates from the Gricean maxim of "Quantity" [11] and is based on simplifying assumptions such as [29]:

² We use the Rhetorical Structure Theory [19] to identify the relations among text segments: according to this theory, text segments participating to a rhetorical relation can have either a more prominent (nucleus) or a less prominent (satellite) role.

Header:	DefineSingleSideEffect (S H ?x ?y)
Domain Constraints:	(Drug ?y) AND (SideEffect ?x ?y)
Intentional Constraints:	NOT (KnowAbout H ?x)
Effects:	(KnowAbout H ?x)
Decomposition:	(Inform S H (Sign ?x)) <i>Nucleus</i>
	(Inform S H (Severity ?x)) <i>Satellite (optional)</i>
	(Inform S H (Frequency ?x)) <i>Satellite (optional)</i>
	(Inform S H (Intensity ?x)) <i>Satellite (optional)</i>
Rhetorical Relation:	ElaborationObjectAttribute

Fig. 3. A typical plan operator, from [7]

1. unlimited working memory of the Hearer (everything an agent knows is always available for reasoning),
2. logical omniscience (every logical conclusion of the Hearer’s knowledge is always available for reasoning) and
3. no autonomy in knowledge (assertions and proposals by the Speaker are accepted by default by the Hearer).

However, as stated by Marilyn Walker [29] and as we saw in our Examples in Sect. 2, it turns out that a topic may be included in a text, even if the Hearer is presumed to already know it, for several reasons: for instance, to augment the evidence in support of a desired belief, to make the topic salient in a given phase of the discourse, or to increase the Hearer acceptance of some claim. The need for mentioning a topic therefore depends, on one side, on characteristics of the Hearers that go beyond their knowledge state, such as their inference ability, their attentional capacity or their emotional state. But it also depends on characteristics of the discourse topic, such as how complex a task is, how hard and how important the inference is, or how acceptable errors consequent to misunderstanding of the message are. When the text has an argumentation purpose (for instance, “to convince the patient to carefully comply with treatment”), the target of the communication process are the Hearers’ goals and intentions rather than, or in addition to, their beliefs and knowledge. The simplifying assumptions leading to avoid repetitions may be rephrased, in this case, as follows:

1. unlimited performance desire of the Hearer (everything the agent intends to do is immediately turned into action),
2. behavioural total coherence (every intention consequent of the Hearer’s mental state is immediately adopted),
3. no autonomy in behaviour (requests are accepted by default by the Hearer).

To account for affective factors in the definition of plan operators’ constraints, conditions under which these assumptions are likely to be violated, and attitudes of the Hearer’s mental state that may affect them, need to be established.

Figure 4 shows an example of how the operator of Fig. 3 may be revised, to plan an affective discourse. This operator is meant to capture a situation in which the considered drug has “serious side effects” and the patient has a high “self-care attitude” and does not fear too much about painful consequences of taking the drug. New items are introduced in the Intentional Constraints slot: the Hearer’s interest towards knowledge of health care implications (in our

Header:	DefineSingleSideEffect (S H ?x ?y)
Domain Constraints:	(Drug ?y) AND (SideEffect ?x ?y) AND (Eq ((Severity ?x) HIGH)
Intentional Constraints:	(Val(KnowAbout H ?x) < t1) AND (Val(WantsToKnow H ?x) > t2) AND (Val(Fear H PAIN) < t3)
Main Effects:	Increases Val(KnowAbout H ?x)
Side Effects:	Decreases Val(Goal H (Take H ?y)) AND Increases Val(Fear H PAIN)
Decomposition:	(Inform S H (Sign ?x)) <i>Nucleus</i> (Inform S H (Severity ?x)) <i>Satellite (optional)</i> (Inform S H (Frequency ?x)) <i>Satellite (optional)</i> (Inform S H (Intensity ?x)) <i>Satellite (optional)</i>
Rhetorical Relation:	ElaborationObjectAttribute

Fig. 4. A plan operator for *affective* discourse planning

case, the side effects of the drug: (WantsToKnow H ?x)) and his/her emotional state. The particular emotion category considered in the example is ‘fear’ about ‘the prospective undesirable event’ of ‘feeling pain due to the side effects of the drug’ [9]. All conditions in this slot are transformed from first order formulae to conditions on the value of the ‘intensity’ of items: the operator will be activated only when this value (the ‘level of knowledge’, the ‘level of interest’, the ‘level of fear’) exceed or are below a threshold value. A different operator would consider the case of patients whose ‘self-care attitude’ is not very high or whose fear about feeling pain is high: in such case, details about most negative aspects of side effects will be hidden, or will eventually be attenuated by appropriate techniques.

Effects. The communication of a topic produces several effects on the Hearer’s mental state: some of them are “main (usually positive) effects” that satisfy the main communicative goal of the Speaker, while others are “side effects” that may be “negative”, not intentional and not avoidable. This makes too simplistic the on/off hypothesis, according to which agents either have or have not a goal, a belief or an attitude, and reverse their state of mind after receiving a communication. Some more fine grained representation is needed, instead, to express the graduality of mental state change.

Let us compare, again, the two operators in Fig. 3 and Fig. 4. In the operator employed in affective discourse planning, the main effect is to increase the Hearer’s knowledge about the side effects of the drug. In addition, a new category of effects is introduced: the “Side Effects” slot (side effects of the communication process!) specifies that the consequences of being informed about the side effects (of drugs!) is a decreased intention of taking the drug (Goal H (Take H ?y)) and an increased fear about their possible negative consequences (Fear H PAIN).

4.2 Sentence Planning and Plan Revision

In the sentence planning step, a number of activities are performed, all having in common the characteristics of producing, from an input discourse plan, a new plan that is, in some sense, “optimized”.

Aggregation, that is the grouping together of some text segments in order to enhance coherency and fluency, is considered the most important of these activities, in the sentence planners that have been proposed so far [5].

Highlighting of primary subjects and reordering of plan tree components are other ‘affect enhancing techniques’ that may be applied, if needed, after a plan has been produced and before surface realization starts.

These sentence planning algorithms should not, however, be driven only by style considerations and should only be applied in specific circumstances, that depend on the Hearer’s affective state on one side, and on the domain topics on the other side.

Let us consider, for instance, aggregation. The decision of whether to aggregate sentences or clauses or whether, on the contrary, to emphasize repetitions in sentences, should consider not only aesthetic parameters, but also the *Hearer characteristics*, the *degree of difficulty* of the topic and its *importance*.

Example:

Let us suppose that the considered drug has three side effects (nausea, headache and insomnia), that each effect has its own severity, frequency and intensity and that the operator in Fig. 4 has been applied in a discourse plan in which side effects are described in sequence. A text with the following content will be produced in this case:

“However, I must inform you that this drug may cause some side effects. The first one is nausea; it is serious, it occurs infrequently, in a strong form, in sensitive patients. The second one is headache; it is serious, it occurs infrequently, in a strong form, in sensitive patients. The third one is insomnia; it is not serious, it occurs frequently, in a strong form, in sensitive patients.”

This text includes several repetitions, some of which are concerned with “positive” aspects of this topic (*infrequently, in particularly sensitive patients*), other with “negative” ones (*it is serious, in a strong form*). To avoid increasing the patient fear about the consequences of taking the drug, a plan revision algorithm may be applied to enhance positive aspects through emphasized repetition and to mitigate negative ones through aggregation. The following revised text will be obtained:

“However, I must inform you that this drug may cause some side effects. A first group of them includes nausea, which occurs infrequently and only in particularly sensitive patients, and headache which, again, occurs infrequently and only in particularly sensitive patients; these side effects are both serious. Then, you may have insomnia: it is not serious but can be frequent; however, once again I would like to reassure you that it occurs only in particularly sensitive patients. All these side effects can occur in a strong form.”

Similar considerations may be applied to the other sentence planning techniques we mentioned before.

4.3 Surface Realization

In the surface realization step, the way that a sentence is rendered (words choice, sentence structure and style employed) depends, as well, on the topic and on the Hearer characteristics. Affective text may be obtained by employing rule-based

heuristics that define when and how empathy elements have to be introduced in the text. Let us see an example of these heuristics, from the previous text:

Example:

To produce the “*however, once again I would like to reassure you that...only...*” from the revised text plan, a rule of the following type is used:

“IF the subtree Explain negative effects of a drug includes “relevant for compliance” items which take unfavourable values AND the patient is particularly fearful about this type of effects, THEN de-emphasize these items”

Similar rules would produce the various *little bit, very, really, may be, kill of* densifier adjectives, adverbs and verbs in the example in Section 2.

5 From Theory to Practice

We have applied some of the principles outlined in the previous Section to our tool for generating explanations about drug prescriptions [6], with the aim of producing more affective messages. We did not revise, so far, the discourse planning component: this would require a theory for stating how *intensity* of knowledge, intentions and emotions should be updated after each operator has been applied, or, in other words, a definition of the Decrease and Increase functions in the two Effect slot of the plan operator in Fig. 4, which is beyond the scope of our present work. We rather focused our efforts on the sentence planning phase, by specifically looking at the three affect-conveying operations encountered in our corpus: treatment of repetitions, highlighting of relevant subjects and reordering of text spans. To this aim, we introduced a plan revision step in the architecture of our generator, after extending the Speaker/Hearer Modelling components and the Domain Knowledge Base (DKB).

The Hearer Model is enriched with new features that formalise interests, preferences and a few emotion types (fear, hope, admiration, reproach and gratitude). Information items in the DKB are labelled according to hypotheses about relationships between the Hearer’s knowledge and those aspects of the Hearer’s mental state that the Speaker (the doctor) intends to achieve. For instance:

- an item ?x whose knowledge may affect the patient’s *capability* to apply correctly the suggested treatment plan for a drug ?y is labelled as ‘relevant-for-correct-treatment’. The cognitive hypothesis, in this case, is that:

(KnowAbout H ?x) → Increases Val (KnowHow H (Take H ?y)).

An example of such an item is ‘drug intake intervals’.

- an item ?x whose knowledge may affect -either positively or negatively- the patient’s *intention* to follow the suggested treatment, for instance to take drug ?y, is labelled as ‘relevant-for-compliance’ and may take ‘favourable’ or ‘unfavourable’ values. The cognitive hypothesis, in this case, is that for favourable items:

(KnowAbout H ?x) → Increases Val (Goal H (Take H ?y))

and for unfavourable items:

(KnowAbout H ?x) → Decreases Val (Goal H (Take H ?y)).

Relevant for compliance items are related to *events* that may affect goal-based emotion categories: for instance, a ‘serious’ or ‘frequent’ side effect of

a drug will increase the risk of ‘feeling pain’, which raises the Hearer’s emotion of ‘fear’. They may be related, as well, to *acts* that affect belief-based emotion categories or compound emotions: for instance, believing that the doctor ‘cares about the patient’s health state’ may affect the patient ‘gratitude’ towards the doctor [9].

In our affective natural language generator, plan revision algorithms are activated by rule-based heuristics that specify when and how affect has to be added in a text. Examples of rules aimed at, respectively, treating repetitions, highlighting relevant subjects and reordering text spans are the following:

- R1:** IF the subtree ‘Explain negative effects of a drug’ includes unfavourable items with the same value, AND the patient is fearful about pain, THEN aggregate these repetitions.
- R2:** IF at least one leaf of the plan tree is a ‘relevant-for-correct-treatment’ item AND the patient is aged THEN substitute the ‘Request to perform treatment’ with a sentence which introduces the most significant of them.
- R3:** IF the side effects mentioned in the plan tree are many AND most of them are serious, THEN make sure that the administration details are mentioned before the side effects.

These rules are represented internally by an algorithm that explores the plan-tree in a top-down way, and finds out, in order, whether reordering, highlighting or treating repetition methods have to be applied.

The activated *plan revision* algorithms apply to the rhetorical structure of the plan tree to produce a new plan tree in which the intended operation is performed. These algorithms exploit several properties of the discourse tree, the main of which is the concept of “most nuclear part” defined by the Rhetorical Structure Theory [19], to guarantee that discourse coherence is preserved after plan revision:

- The module for *treating repetitions* classifies the leaves in the subtree of interest according to their role (nucleus or satellite) and to their value of ‘relevance-for-compliance’: a new subtree is then produced, in which topics with ‘unfavourable’ characteristics are grouped, repetitions of these characteristics are pruned out and the correct nuclearity or the discourse is restored.
- The module for *highlighting relevant subjects*, on the contrary, identifies ‘relevant-for-correct-treatment’ topics among the plan tree leaves, in order to produce the minimal subtree that generates a summary sentence in which these topics are mentioned. Such a subtree is subsequently grafted into the original plan tree, in an appropriate position.
- Finally, the module for text spans’ reordering applies properties of exchangeability of satellites in mononuclear RRs and of nuclei in some multinuclear RRs, to exchange the relative position of two text spans, by again insuring a final, coherent tree.

Figure 5 shows an example of highlighting subtree (on the right) that was extracted from the the part of the discourse plan that instructs on how to take a specific drug (on the left). The ‘relevant-for-correct-treatment’ topics are, in this example, the drug administration modalities (*by mouth*, *with water*); a ‘relevant-for-compliance’ item (drug efficacy) is also included in the minimal subtree, to

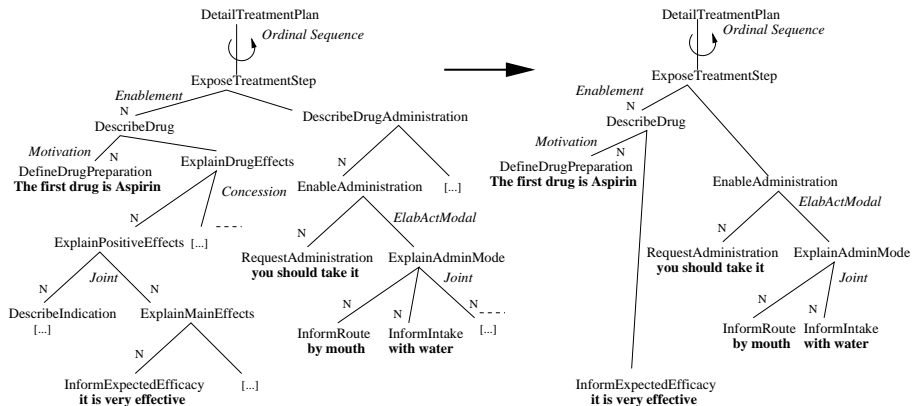


Fig. 5. Extraction of an highlighting subtree

increase the patient’s intention to take this drug (as in segment 2 of Doctor 2’s text, in Fig. 1).

The plan revision algorithms, the hypotheses on which they are based and the experiments which led to their design are described in detail in [6]. Extension of the affect-conveying techniques to the other phases of natural language generation (planning and surface realization) is a work still in progress.

6 Conclusions

In this paper we argued for the need for more research about how affect can be introduced in the automated natural language generation of explanatory and advisory texts, and we sketched some ideas that may be useful in this respect.

Far from advocating a flouting of Grice’s maxims about natural language communication [11], we believe that Affective Text Generation requires interpreting Gricean maxims in a “relaxed” and extended way. When Grice suggests to “make your contribution as informative as required”, this should not be interpreted as “never repeat something the user already knows, desires or may infer”, but as “avoid repeating more than what is really needed”. When he recommends to avoid “ambiguity” or “obscurity of expression”, this should not be interpreted as “be always fully sincere”, but as “avoid insincerity while not needed”, ... and so on. In this perspective, Grice maxims may be seen as default rules, and affective texts as exceptions in their application.

As a final observation, our discussion concerned how the user’s emotions can influence a computer system producing text. A completely different issue concerns whether the computer system should show *its own* “emotions” in generating natural language. Studies have proven that, although conscious of being interacting with a machine, humans have a clear perception of the computer system’s *personality* [20,8]. On the other hand, it is still to be established whether users would appreciate a computer system showing the same level of empathy a

human being would show, and in fact studies produced some evidence that this may not always be the case [4].

References

1. H. M. Becker, editor. *The Health Belief Model and Personal Health Behavior*. Thorofare, N.J: C. B. Slack, 1974. 204
2. D. C. Berry, T. Gillie, and S. Banbury. What Do Patients Want to Know: an Empirical Approach to Explanation Generation and Validation. *Expert Systems With Applications*, 8(4):419–428, 1995. 205
3. C. Castelfranchi, F. de Rosis, and F. Grasso. Deception and Suspect in Medical Interactions: Towards a simulation of believable dialogues. In Y. Wilks, editor, *Machine Conversations*, volume 511 of *International Series in Engineering and Computer Science*, chapter 8. Kluwer, 1999. 207
4. C. Cheepen and J. Monaghan. Naturalness in Automated Dialogues - less is more. In D. Levy and Y. Wilks, editors, *First International Workshop on Human-Computer Conversation*, pages 83–89, 1997. 217
5. H. Dalianis. *Concise Natural Language Generation from Formal Specifications*. PhD thesis, Royal Institute of Technology/Stockholm University, Department of Computer and System Science, June 1996. 213
6. F. de Rosis, F. Grasso, and D. C. Berry. Refining Instructional Text Generation After Evaluation. *Artificial Intelligence in Medicine*, 17(1):1–36, 1999. 208, 214, 216
7. F. de Rosis, F. Grasso, D. C. Berry, and T. Gillie. Mediating Hearer’s and Speaker’s Views in the Generation of Adaptive Explanations. *Expert Systems With Applications*, 8(4):429–443, 1995. 210, 211
8. D. C. Dryer. Getting Personal with Computers: How to design personalities for agents. *Applied Artificial Intelligence*, 13(3):273–295, 1999. 216
9. C. D. Elliott, J. Rickel, and J. Lester. Lifelike Pedagogical Agents and Affective Computing: An Exploratory Synthesis. In M. Wooldridge and M. Veloso, editors, *Artificial Intelligence Today*, number 1600 in *Lecture Notes in Computer Science*, pages 195–212. Springer-Verlag, 1999. 212, 215
10. F. Grasso, A. Cawsey, and R. Jones. Dialectical Argumentation to Solve Conflicts in Advice Giving: a case study in the promotion of healthy nutrition. *International Journal of Human-Computer Studies*, (to appear). 207
11. H. P. Grice. Logic and Conversation. In P. Cole and J. L. Morgan, editors, *Speech Acts*, volume 3 of *Syntax and Semantics*, pages 41–58. New York: Seminar Press, 1975. 210, 216
12. I. Haimowitz. Modeling all Dialogue System Participants to generate Empathetic Responses. *Computer Methods and Programs in Biomedicine*, 35:321–330, 1991. 207
13. E. Hovy. *Generating Natural Language under Pragmatic Constraints*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1988. 208
14. E. Hovy and L. Wanner. Managing Sentence Planning Requirements. In K. Jokinen, M. Maybury, M. Zock, and I. Zukerman, editors, *Proceedings of the ECAI-96 Workshop on: Gaps and Bridges: New directions in Planning and NLG*, pages 53–58, 1996. 208
15. A. Jameson and T. Weis. How to Juggle Discourse Obligations. Technical Report 133, University of Saarbrücken, Department of Computer Science, October 1996. 207

16. M. E. Kölln. Employing User Attitudes in Text Planning. In *Proceedings of the 5th European Workshop on Natural Language Generation*, pages 163–179, May 1995. 208
17. J. Lewi and B. Hayes-Roth, editors. *Proceedings of the 1st International Conference on Autonomous Agents (Agent97)*. ACM Press, February 1997. 218
18. A. B. Loyall and J. Bates. Personality-Rich Believable Agents that Use Language. In Lewi and Hayes-Roth [17], pages 106–113. 208
19. W. Mann and S. Thompson. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text*, 8(3):243–281, 1988. 210, 215
20. Y. Moon and C. Nass. How “Real” Are Computer Personalities? Psychological Responses to Personality Types in Human-Computer Interaction. *Communication Research*, 23(6):651–674, 1996. 216
21. J. Moore. *Participating in Explanatory Dialogues*. MIT Press, Cambridge (Mass.), 1995. 209
22. R. W. Picard. Does HAL Cry Digital Tears?: Emotion and Computers. In D. Stork, editor, *HAL’s Legacy - 2001’s Computer as Dream and Reality*, chapter 13, pages 279–303. MIT Press, Cambridge (Mass.), 1996. 204
23. M. Reape and C. Mellish. Just what is aggregation anyway? In *Proceedings of the 7th European Workshop on Natural Language Generation*, pages 20–29, 1999. 208
24. E. Reiter and R. Dale. Building Applied Natural Language Generation Systems. *Natural Language Engineering*, 3(1):57–87, 1997. 204, 209
25. J. A. A. Sillince and R. H. Minors. What Makes a Strong Argument?: emotions, highly placed values, and role playing. *Communication and Cognition*, 24(3/4):281–298, 1991. 207
26. W. Stroebe and M. Stroebe. *Social Psychology and Health*. Mapping Social Psychology. Open University Press, Buckingham, 1995. 204
27. K. Sycara. Persuasive Argumentation in Negotiation. *Theory and Decision*, 28(3):203–242, 1990. 207
28. M. A. Walker. Redundancy in Collaborative Dialogue. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING92)*, 1992. 208
29. M. A. Walker. The Effect of Resource Limits and Task Complexity on Collaborative Planning in Dialogue. *Artificial Intelligence*, 85(1-2):181–243, 1996. 208, 210, 211
30. M. A. Walker, J. Cahn, and S. Whittaker. Improvising Linguistic Style: Social and Affective Bases for Agent Personality. In Lewi and Hayes-Roth [17], pages 96–105. 207

An Interview with Rosalind Picard, Author of “Affective Computing”

Q: Can you tell a little bit about the area of “Affective Computing” and the reasons and motivations that drove you to explore this difficult and at the same time controversial theme?

A: I started thinking about emotion in 1993 while I was trying to give computers better perceptual abilities. I had been developing algorithms for machine vision and for modelling video content, and I was interested in extending the machine’s visual capabilities to other perceptual domains, as well as enhancing its visual abilities. Over a vacation, I read one of Cytowic’s books about synesthesia – an interesting phenomenon in humans where the perceptual senses appear to interact in an unusual way, a phenomenon that pointed to limbic system activity in the brain. All the human vision literature I had read had focused on cortical structures, not limbic, so I became curious as to what might be going on beneath the cortex that had to do with vision. I began to read more neuroscience, where I gradually began to see that emotion, which has its home in the limbic system, not only plays a role in human perception, but in many other human aspects of intelligence that we were trying to emulate in machines. It struck me that I had never heard AI researchers address the role of emotion in perception, decision-making, memory retrieval, creativity, and more, and yet it seemed really important. Furthermore, some of the more recent neuroscience findings showed that people whose emotions were impaired behaved in a less intelligent way, a way that I recognised as having problems similar to AI rule-based systems. I found a lot of ideas that I thought AI researchers should know about, unaware at the time that there had been any work in AI on emotion. Later, while digging through the literature and talking with colleagues, I learned that several researchers had not only thought about giving machine emotions, but had built preliminary systems along these lines. Most of these were not well known, and were very limited in their modelling of emotions. As I dug more, reading about the work that had been done, I was struck that there still seemed to be very little understanding among computer scientists about the powerful beneficial roles that emotions appear to play in humans, about the complexity of the human emotional system, and about the potential for imitating some of these functions in machines.

In the meantime, I had a fully-funded research program in video content based retrieval, and it was not easy to switch my research agenda. Fortunately, I am in a lab that encourages risk and change, and several colleagues and sponsors of my previous work showed interest in my investigating this “crazy and impossible-sounding” area. In writing a thought-piece in January 1995 and a book in 1997, I tried to build a framework for what I called *affective computing*, to bring together researchers from neuroscience, computer science, psychology, cognitive science, pattern recognition, signal processing, medicine, and the social

sciences to begin to address this difficult area in a thorough and innovative new way.

Q: Some part of the research of your group goes into capturing emotional states of the user, from the sensing to the recognition of emotions. Can you briefly describe what processes and techniques would be needed for example, for my computer to be able of recognise my emotions at the present moment?

A: We Still don't know exactly what would be needed in the minimal sense of this word; it is too early to tell, and it may differ for different people. The issue is complicated by many factors, perhaps the biggest factor being that individuals have vastly different comfort levels with different kinds of sensors, and I think the human needs and desires should come first. Some people are happy to have a camera pointed at their face and to show genuine facial expressions to the system, while others consider a camera a major privacy invasion. Some computers are being worn in clothing, and this affords different kinds of sensors that can be even more comfortable than cameras. Our research is examining many different situations – from traditional desktop (keyboard/mouse only) sensing to innovative wearable computers that afford sensing of physiological and other behavioural signals.

Q: Some of the results achieved in your team show that you can recognize with a success rate of about 81%, 8 emotional states (Neutral, Anger, Hate, Grief, Platonic Love, Romantic Love, Joy and Reverence), using a physiological based recognition with an actress expressing those emotions over a period of time. Can you tell us a little bit more about these impressive results and what type of signals you are handling? Why did you focus on those 8 emotional states?

A: These results are on data taken from one person over an extended period of time (twenty days of collection, spanning about five weeks), in contrast with efforts in psychophysiology that usually focus on lots of subjects over a short amount of time (usually less than an hour). Our interest was on seeing if physiology would show recognisably characteristic patterns for the emotions day in day out. Previous results on large subject pools have shown a few highly-averaged characteristics that correspond with certain states, but the results were not useful for recognising emotional information for an individual, who may vary from this average. The signals we used were: blood volume pulse, electromyogram (muscle tension), skin conductivity, and respiration. We chose these signals initially because we had an FDA-approved sensing device that seemed fairly reliable for these signals and these signals were believed to show some changes with emotion. This particular set of eight emotions is an unusual one to start with, but we chose it because we had a copy of Clynes' sentograph with related software, which has been used around the world to help subjects actually feel the set of emotions in each session. The choice, ordering, and means of triggering each of the emotions has been fine-tuned by Clynes and, in our experience, was a significant aid to the subject in actually creating the feeling of each emotion day in-day out. Since we

were especially interested in recognising felt emotions (vs. just a communication of an emotion, as if an actor was onstage) it was important that we made sure our subject could self-generate the emotional feeling. (Note that Clynes, while not a mainstream emotion theorist, is a very fine musician, and chose his list largely to include emotions communicated in music; hence we find emotions like love and reverence, which are not on most theorists' lists.)

Q: How do such results relate with other affective states such as mood?

A: We really haven't worked on mood recognition yet, and I don't know of any results in this area, so I cannot say.

Q: But don't these physiological signals depend very much on the person, the age, the gender, (the mood), and especially on his/her hormone levels?

A: There is some evidence for person-dependent patterns, much like each person's voice is different. But there have also been some interesting consistencies found across people. Our interest is in starting with person-dependent recognition, and then grouping similar persons to extend the results to person-independence (similar to the successful strategy used by speech recognition researchers.) The experiment above was the first of its kind in looking at one person over more than a month where hormone variations, differences in daily caffeine and sugar consumption, and different moods would be present. Whether or not we could obtain highly significant recognition accuracy under these variable conditions was a big question. We were delighted to find that we could, although we also found that day-to-day variations were significant enough that the formulation of a daily baseline was critical.

Q: Other than physiological signals, recognising user emotions through facial expression or through voice analysis seems to also lead to quite impressive results. How does the success of these recognition modalities compare with the success of the physiological recognition?

A: So far there have been no comparative studies that consider one subject over a long period of time, where the subject genuinely tried to feel each emotion. Much of the data collected for facial expression or vocal expression recognition is exaggerated or gathered from actors trying to express and not necessarily feel the emotion, and it is hard to tell how much that has biased the results. It is also hard to tell how genuine the self-induced emotions in our subject are; she often indicated strongly feeling them, being moved to tears during grief, and so forth—but also admitted that she had some difficulty self-inducing anger to the same intensity that it could occur in real life, and that this was dependent on her mood for the day. In general, automated recognition results of six categories of affect in speech tend to be lower than the 81% accuracy we obtained based on eight categories of emotion with physiology, which in turn is lower than the

80-98% accuracy obtained on a set of six facial expressions (typically the Ekman six.) It should be noted that rates vary under different experimental conditions, and the conditions have not been held constant across these domains. Also, recognition of arousal in speech and in physiology seems to be very good while recognition of valence from facial expressions is very good, so that combinations of modalities should give more robust results. Note that all the results reported to date are on forced-choice selection of a small number of emotion categories, on pre-segmented data, much like in the early days of speech recognition where the computer tried to recognise which of eight digits, 1à8, the user was speaking. Continuous detection and recognition of emotion is a harder problem.

Q: According to Kaiser & Wehrle (in this book) emotion recognition needs to be done within a temporal and situational context. Do you think that emotion recognition should be enriched with some kind of “environment” analysis in order to capture at least some aspects of the situational context of the user?

A: Absolutely. We have projects where we analyse the situation jointly with expression, e.g., is the user repeatedly clicking with signs of increasing frustration or with a relaxed curious state? In my book I gave an example of a person having the winning lottery number where reasoning about the situation (they should be happy) does not agree with the expression on their face (panic, despair). When situational and expressive variables do not match, this leads to an interesting state. When they do match, their agreement can be used to boost confidence in the analysis.

Q: Emotion recognition is only necessary if there is a clear need for a computer/machine to react differently to one or the other emotion, whenever detected, and for the benefit of the user. Can you give us an idea of situations where such adaptation is necessary?

A: I’m not sure I agree with the premise of the opening statement—if I have previously said that, then I would word it more carefully today. Recognition is never “necessary” if users don’t want it – no matter how beneficial designers might think it is for them. I think it is essential that we honour the users’ feelings first in any system design. That said, let me give an example where adaptation might be useful and appreciated: My students and I have built a number of systems that “learn” based on feedback from the user, and to date this feedback has forced the user to stop what he or she is doing, explicitly rank their opinion, and communicate this via keyboard and mouse to the computer. This is time consuming and often taxing on the user – it would be much more natural and effortless if the computer just watched for signs of pleasure or displeasure, interest or boredom, etc. while you used the system, and occasionally asked for clarification or explicit feedback. Since learning systems are being integrated into a variety of tools these days – computer search algorithms, shopping bots,

devices that interrupt you with information, and more, many users would like to find ways to make interaction with them more comfortable and natural.

Emotion recognition might also be useful in some systems that aren't meant to react differently to what the user is expressing. We have developed one system that “recognised” direct negative or positive input from the user not because there was a clear need for the system to react differently, but because we hoped the information could be meaningful to the designer of the system in learning about its interaction with users in an ongoing way.

Q: In his most recent book, Damásio discussed that the problem of “feeling” emotions is based on the notion of “self” and therefore, there cannot be a “feeling of an emotion” without consciousness. Does it mean that the work on affective computing will be forced to consider the long and extensively debated problem of consciousness?

A: I agree with Damásio on this issue. Our having of emotional feelings is what many of us refer to as the issue of “emotional experience”. Some aspects of consciousness seem to be required for emotional experience. Will this hold up research in affective computing? I don't think so, because I think that emotional experience is only one of many aspects of giving machines emotions, and the latter is only one part of affective computing research. A huge amount of progress toward affective computing can be made without worrying about consciousness. However, I find it interesting to try to stay abreast of research on consciousness, because it will play a critical role in understanding if we can build machines with emotional experience that approximates ours. Unlike many of my colleagues who think the latter is merely a matter of time, I do not see it as a foregone conclusion at all; I have yet to see any scientific evidence that the kind of emotional experience we have can be duplicated in a machine.

Q: Some of the work on emotion synthesis is becoming extremely important in the construction of human-computer interfaces, in particular in computer characters that express emotions, show empathy with the user, entertain the user etc. In this book Paola Rizzo argues that in order to achieve affective interactions such characters do not necessarily need to “have” emotions. Further, Reeves and Nass have shown that people in general ascribe human properties to the interfaces and treat computers in a human-like manner. One may ask then, to what degree one should explore emotion synthesis if one can achieve “affective interactions” without it anyway?

A: I agree with Rizzo that it is not necessary. Animators for years have done a powerful job of crafting emotion expression by hand, with no generative models. However, animators today are recognising the power of automating some of this process with generative models, if only to offload some of the more tedious aspects of their work. But I think synthesis can go much further than this: emotion synthesis should not just influence expression and behaviour, but also internal

workings of the synthetic character, things that may not be seen immediately to those interacting with the character. The emotion system could work behind the scenes, guiding what the character attends to and how that information is learned, stored, and retrieved. Synthetic emotions could guide decision-making, perception, and motivation, as well as visible expression and behaviour. And, building such systems will raise important questions about how human and animal emotions work – no doubt inspiring new research in living systems.

Q: In your own terms, affective computing is computing that “relates to, arises from or deliberately influences emotions”. Does it mean that designing user interfaces and computer applications to influence users’ emotions, such as computer games in general can be considered affective computing? For example, during the workshop Tetris was mentioned as good example of a game with a strong emotion arousal leading to a catharsis-like explosive relief of anger.

A: There is a grey area in my definition, like in many definitions. People can have emotional reactions to just about anything, and to call all such things “affective” would soon render the definition meaningless. When I inserted the word “deliberately” before “influences” what I had in mind was that the designer’s considerations of the emotional impact influenced his or her design. With many designs it is tricky for the designer to articulate their thought process. Many designers “feel” their way through the design, and are strongly influenced by its emotional characteristics, but do not systematically or consciously attend to emotional characteristics; yet that is what their intuition is doing. In these cases perhaps it is an example of affective computing, but I don’t particularly mind whether it is called that or not.

Out of curiosity and to check my thinking, I posed this question to one of the designers of Tetris, Vadim Gerasimov. He said, “I don’t think Tetris was specifically designed to be affective. In fact we never really discussed the emotional aspects of the games we designed before they were ready. We were looking for random ideas of engaging game environments never defining what engaging meant.” He added, “I’ve never experienced strong emotional arousal nor explosive relief of anger with Tetris. So it may not be the best example of affective computing. On the other hand, some people take success and failure in the puzzles far more seriously. ..Tetris, as many other computer games, always ends with a failure. The objective of the game is to avoid the failure as long as possible. When the game ends, the player always has to seek a relief of the anger somewhere outside of the game.”

Q: Doesn’t this “manipulation of emotions” lead to some difficult ethical problems related with individual freedom?

A: Manipulation of emotions leads to ethical questions, whether it’s one person directly manipulating another person’s emotions, or indirectly, via a machine.

When we manipulate our own emotions (via choosing music, or choosing to play a video game that will manipulate them for us) then manipulation is usually seen as good. I don't think the ethical questions are substantially different for affective computing than they are when a computer is not involved, but I do think it is an area we should take interest in educating people about. Some people "trust" things more when they come from a computer, and this could be problematic.

Q: However, when we think of E-Business we think of "trust", and it seems that the presence of affect in E-business interactions may have a negative impact, and users may consider such applications "non-trustworthy". Do you think there is such a danger?

A: If you believe the theory of Nass-and-Reeves, then it will be perceived as trustworthy if it behaves in a way that is like a trustworthy person behaves. It is known that people trust a message from another person more if all their modes of expression agree (the message is consistent with their voice, face, gestures, etc.) and trust them less if these modes disagree. The computer may have to mimic this, or err on the side of showing no expression, since it really can't mimic the human body. If the affective abilities are appropriate to the business situation (read: minimal expression, savvy about signs of liking and disliking and respectful handling of this information) then these abilities should contribute to the success of the interaction. On the other hand, if "affect" means "some character showing a lot of emotion in an insensitive way" then I think it's been wrongly implemented, and will be detrimental to the interaction.

Q: Part of your research involves the construction of "affective toys" that not only capture some affective states of the children but also provide affective responses. Can you tell a little bit more about this area?

A: We have done two projects in this area. For the first project, Dana Kirsch and I designed and built a "tigger" that has five internal states, each corresponding to a pattern of human emotional behaviour. The toy also has a half dozen sensors for detecting things like if it is being postured upright and bounced. The detection of this behaviour, for example, leads to a happy internal state where tigger's ears move upwards and it makes a happy-sounding vocalisation. The idea is that the toy can serve as an affective mirror, showing a reflection of the affective qualities with which the child is playing. We evaluated the toy with kids and found that it can successfully communicate a number of emotional qualities.

Q: Some of the experiments with the affective toys were with autistic children. How were the results obtained? Do you believe that affective computing will play an important role for the communication with children and people with social, motor and mental difficulties?

A: Our second project involving toys was the design and construction of a system to help autistic kids. This project, Kathi Blocher's masters thesis, "ASQ" for Affective Social Quotient, used stuffed animals (four dwarfs – which signified angry, sad, happy, and surprise), as a wireless interface to a computer that played different emotional scenarios. Gathering this footage was half the battle – finding footage that would capture and hold autistic kid's interest, while including a broad range of examples of the four emotions. Six kids participated in trials using the system at the Dan Marino Centre in Miami, where they were rewarded for choosing the toy consistent with the emotional situation played by the computer. Several of the kids showed improvement over a few weeks of sessions in accurately choosing the right emotions. One child demonstrated generalisation in a home setting, which was very exciting.

I believe that the problems we face in giving computers affective skills are very similar to the problems people who work with autistics face. As we develop tools for computers, we should extend them to autistics, since many of them may potentially benefit from this technology.

Q: Do you think that one day we will be able to build a machine, which like HAL, is able to express, "understand", and perhaps "feel" emotions?

A: I don't have a short answer to this, because of the complexity involved in the notions of "understanding" and "feeling" emotions, and because of the difference between giving the appearance of these things, and really having them at a non-trivial level. The big problem of course is the aspect of consciousness that permits awareness of feeling in the same way that we people think of it. To read more about "building a HAL with emotions" let me refer people to the chapter I wrote in the book *Hal's Legacy*, edited by David Stork, MIT Press, 1997.

Q: Do you think that affective computing may change the gender differences we find nowadays in the areas of computing?

A: The primary gender-related difference I know about affect is that women as a group score better than men score as a group when it comes to recognising emotion. However, women in computing may not score better; it is an interesting subset to compare, and I don't know of any such studies within our field. The technical fields have traditionally attracted people with a higher likelihood of having autism in their family history, as compared to the non-technical fields, further adding ammunition to those who think of computer scientists and engineers as not very good at things related to feelings. By recognising important roles of emotion, and trying to build these into machines, we might not only help educate current computer scientists about emotional skills, but perhaps it will attract new kinds of computer scientists. In my own experience, I've never had such a high percentage of women in my research group as I've had since including affect on the agenda, but this is only one data point. I would certainly like to see more women in the field, and I think most men would agree with this.

Q: How do you see the future of this fascinating area of affective computing?

A: My hope is that this research area will become one with the highest intellectual and ethical standards, creatively developing radically better interfaces and computing systems for all kinds of people, not just for the computer elite or those in power, but for people of all levels of ability. These systems should always show respect for the user’s feelings, in choice of design and in choice of system responses to each individual. It is a tragedy when affective sensing or goofy affective expressions are imposed on users in forms that annoy or belittle them. I expect that research on constructing high-quality affective systems will contribute greatly to fundamental new understanding about emotion and its interaction with cognition, perception, health, social interaction, and intelligence.

Index

- Affect disorders
 - appraisal and, 58
- Affective behaviour, 139
- Affective computing, 14, 49, 219
 - *definition*, 2
 - discussion of definition, 224
- Affective interactions
 - *discussion*, 2–6
 - in teaching and learning situations, 23–33
 - and children, 35–46
 - animated lifelike characters and, 150–163
 - aspects of emotion for, 98–101
 - entertainment and, 166–178
 - facial expression in, 49–60, 182–193
 - social nature, 173
 - user modeling and, 64–74
- Affective reasoning, 132
 - in the Presence project, 161
- Affective Social Quotient, 226
- Affective states
 - *versus* emotional states, 2, 151
 - attribution, 175
 - duration, 151
 - expression, 151
 - focus, 151
 - in CMattie, 108
 - in IDA, 117
 - intensity, 151
 - representation in user models, 68
- Affective text generation, 204–217
 - discourse planning in, 209–212
 - sentence planning in, 212–214
- Affective toys, 225
- Affective user modeling, 64–74
- AgentLink, I
- Alarm system, 116
- Allen, Steve, 150
- Anderson, A, 15
- André, Elisabeth, 5, 150
- Anger, 3, 38, 57, 220
 - in computer game interactions, 57
- Animated lifelike characters, 5
 - believability and, 6
 - challenges of creating, 196
 - emotion expression in, 6
 - for entertainment, 167–178
 - with Microsoft Agent, 156
 - with personality, 150–163, 197
- Anthropomorphism, 172
- Antunes, Luis, 4, 5, 121
- Appraisal, 91, 94–98
 - in Geneva Appraisal Theory Environment (GATE), 57
 - evaluation and, 96–98
 - facial expression and, 57
 - simulation of, 73
 - structure in the OCC model, 70
- Aristotelian theory, 139
- Aristotle, 98
- Artificial emotions, *see* Synthetic emotions
- Aspy, D., 26
- Attention
 - user attention, 182
- Autonomous Agent Modeling Environment (AAME), 57
- Autonomy, 123
 - belief, 124
 - goals, 124
 - principles for, 124–125
- Ball, Gene, 6, 196
- Bargh, J., 97
- Basic emotions, 41, 51, 53
 - and children, 38
- Bates, J., 172, 175
- Bayesian networks, 200–201
- Behaviour
 - and agent’s internal states, 199
 - emotions and, 111, 139
 - planning, *see also* Decision making
 - in the Virtual Puppet Theater Architecture, 154
- Belief, Desire, Intention (BDI)
 - architecture, 90, 127
 - model, 122
- Beliefs, 4, 91
 - attributional, 78

- categorization of, 78
- emotion and, 78
- for envy, 81–82
- for shame, 82–83
- in emotion recognition, 37
- model of action and communication, 183
- trust and, 83–84
- types of, 84
- Believability, 6, 167
- Believable agents, 23, 167
- BGP-MS, 68
- Blood volume pressure (BVP), 4, 11–14
- Bogner, Myles, 5, 107
- Bozinovski, Stevo, 5, 138
- Breese, Jack, 6, 196
- Brentano F., 139
- Brna, Paul, 21
- Buster Keaton, 175
- BVG architecture, 123, 125–127, 134
 - choice and, 126
- Calhoun, C., 139
- Castelfranchi, Cristiano, 4, 76, 124, 128, 134
- Cathexis, 118
- Central nervous system (CNS), 11
- Children
 - affective interactions and, 35–46
 - and basic emotions recognition, 38
 - Emotion communication and, 6
- Choice, 122
 - in decision making, 122
 - mechanisms, 129–131
- Clore, G., 65, 173
- Clynes, M., 220, 221
- CMattie, 107–112
 - emotional behaviour, 111
 - emotions in, 108
- Codelet, 108
 - emotion, 109
- Coelho, Helder, 5, 121
- Cognition
 - emotion and, 107
- Cognitive Evaluations, 92
- Cognitive Units, 184–186
 - of approving, 185
 - of imploring, 184
 - of praising, 185
 - of warning, 185
- Coleridge, S., 170
- Collins, A., 65, 173
- Communicative acts
 - and facial expression, 184
 - meaning of, 183
 - signal of, 183
- Computer Games
 - interactions in, 57
 - *Pacman*, 176
 - *Tetris*, 224
- Computer Integrated Classroom- CiC, 22
- Computer mediated interactions
 - empathy and, 23
 - measure of quality in, 9–18
- Consciousness, 114, 116
 - and emotional experience, 223
 - codelets, 114
 - in software agents, 107
- Contempt, 38
- Cooper, Bridget, 2, 21
- Coping, 50
- Costa, P., 198
- Crossbar Adaptive Array (CAA), 141
- Cytowic, R., 219
- Damásio, A., 5, 97, 223
- Darwin, C., 52
- Davis, K., 174
- de Rosi, Fiorella, 6, 204
- Decision making
 - emotion and, 121
- Descamps, P., 10
- Desires
 - in emotion recognition, 37
- Dialogue support, 25
- Disgust, 38
- Disney animation, 172
- Disposition, 90
- Drama, 170
- Drives, 91
- Dynamics
 - of facial information, 39
 - of emotional episodes, 50
- Ekman, P., 38, 52, 53, 184, 187, 188, 222
- ELIZA, 172
- Elliott, C., 65
- Ellsworth, P., 38
- Emotion
 - action selection and, 108

- as evaluation, 139
 - believability and, 170–174
 - codelet, 109
 - cognition and, 107
 - cognitive evaluation and, 4
 - conative aspect of, 90
 - decay rate, 110
 - decision making and, 121
 - elicitation, 170, 175
 - entertainment and, 170
 - evaluation and, 94
 - expressions of, 182
 - exaggeration, 40
 - facial expressions and, *see* Facial expression
 - feeling, 88
 - goals and, 79–80
 - in decision making, 132
 - in IDA, 115–117
 - in the Inhabited Market Place, 157–158
 - in the Presence project, 160–162
 - intention and, 77
 - interaction and, 99–101
 - intuitive affective appraisal and, 4
 - learning based on, 138
 - manifestation
 - in agents, 175
 - motivation and, 80
 - motivational aspect of, 90
 - natural language generation and, 6
 - personality and, 154–155, 170, 174
 - processing, 170
 - recognition, 2–4
 - speech and, 3
 - synthesis, 4
 - value or valence, 77
 - without feeling, 88–91
- Emotion pathology
- appraisal and, 58
- Emotion processing
- emotion manifestation and, 176
- Emotion recognition, 220–223
- by children, 37–46
 - caricature faces *versus* normal faces, 39
 - through facial expressions, 49–60
- Emotion synthesis, 223
- Emotion theories
- appraisal, 52
 - componential appraisal theory, 54
 - discrete, 52
- Emotion-based learning, 138
- Emotional agents, 166
- Emotional experience, 223
- Emotional intelligence, 153
- Emotional interfaces, 50
- Emotional learning architecture, 141
- Emotional problem solving, 50
- Emotional states
- bad, 108
 - good, 108
 - in CMattie, 109–112
- Emotionally involved, 172
- Empathy, 2
- attitudes, 28
 - body gestures and, 23
 - body language, 28
 - characteristics, 23
 - content of teaching, 29
 - facial characteristics, 28
 - facial expression, 23
 - in classrooms, 27
 - in intelligent systems, 27
 - in teaching and learning, 21
 - method of teaching, 29
 - motivation and, 23
 - other features, 29
 - positioning, 28
 - positive affirmation and, 23
 - responses, 29
 - voice, 28
 - voice tone and, 23
- Engagement, 168
- Entertainment, 166–178
- Envy, 80–88
- beliefs and goals for, 81–82
- ETNA Project, 16
- Evaluation, 4, 92–94
- appraisal, 54, 96–98
 - emotion, 77, 139
 - emotion and, 94
 - positive and negative, 78
- Expectations, 91
- Facial Action Coding System (FACS), 3, 52, 54, 184
- Action Units, 184
 - in Facial Expression Analysis Tool, 56
- Facial Action Composing Environment (FACE), 53, 57
- Facial expression, 182–193

- appraisal and, 57
- as communicative signal, 51
- as indicator of emotion, 51
- automatic interpretation, 49
- dynamics of, 53
- emotional communication and, 99
- emotional meaning, 182–193
- empathy and, 23
- functions of, 49, 51–52
- intensity of affect in, 39
- raising eyebrows, 186
- recognition by children, 37–46
- synthesis of, 49
- veracity of, 38
- Facial Expression Analysis Tool (FEAT), 56
- Fear, 38, 89
- Feeling, 77, 223
- Five Factor Model (FFM), 151–152
- Franklin, Stan, 5, 107
- Friesen, W., 38, 184
- Frijda, N., 88
- Frown
 - meanings of, 51
- Galvanic skin response (GSR), 4, 11–14
- Gebhard, Patrick, 150
- Gender differences in computing, 226
- Geneva Appraisal Manipulation Environment (GAME), 3, 55
- Geneva Appraisal Theory Environment (GATE), 57
- George, Pat, 6, 35
- Gerasimov, V., 224
- Gestures
 - automatic interpretation of, 49
 - empathy and body, 23
 - personality and, 201
- Global workspace theory, 114
- Gnepp, P., 38
- Goals, 4, 91
 - adoption, 5, 125, 128, 131
 - emotion and, 79–80
 - generation, 127, 128
 - model of action and communication, 183
 - selection, 127
- Goldberg, A., 198
- Grant, J., 67
- Grasso, Floriana, 6, 204
- Grice's maxims, 210, 216
- Grice, H., 210, 216
- Grief, 3, 220
- HAL, 226
- Happiness, 38
- Hate, 3, 220
- Hearer's model, *see also* User model, 204–205
 - for affective text generation, 214–216
- Heart rate (HR), 4, 11–14
- Heider, F, 172
- Hormone variation, 221
- Hovy, E., 208
- Human-Agent interaction, 155
 - emotions in, 98
- Human-computer interaction
 - facial expression, 55
- Humor, 175
- Hunger, 90
- i3 Net, I, 21
- IDA, 107–108
- Improv, 198
- Intelligent agents, 138
 - architecture in the Virtual Puppet Theater, 154
 - autonomy of, 123–125, 127
 - with values, 122
- Intelligent Distribution Agent (IDA), 113–117
- Intention
 - emotion and, 77
- Inter-Agent communication
 - emotions in, 98
- Interactive Data Elicitation and Analysis tool (IDEA), 56
- Interest, 38
- International Telecommunications Union (ITU), 10
- Intuitive Appraisal, 92
 - cognitive evaluation, 92
- Izard, C., 52
- Jones, E., 174
- Joy, 3, 220
- Kölln, M., 208
- Kaiser, Susanne, 3, 49, 54, 222
- Kirsch, D., 225

- Klesen, Martin, 150
 Knoche, H., 11
- Laurel, B., 168, 170
 Learning agent, 140
 - emotion based, 140
 - reinforcement based, 140
 - tutorial based, 140
- LeDoux, J., 133
 Leventhal, H., 50
 Lifelike characters, 150, 167
 Limbic system, 116, 219
 Love, 3
 - Platonic, 220
 - romantic, 220
- Loyall, B., 172
- Machado, Isabel, 64
 Manipulation of emotions, 224
 - ethics and, 225
- Mann, W., 193
 Markham, R., 38
 Martinho, Carlos, 64
 Martins, Alex, 21
 Matthiessen, C, 193
 McCauley, Lee, 5, 107
 McCrae, R., 198
 McIlhagga, Malcolm, 6, 35
 Mean Opinion Score (MOS), 10
 Mellish, C., 208
 Method of teaching
 - empathy, 29
- Microsoft Agent, 156
 Minsky, M., 108
 - Society of Mind, 108
- Mistrust, 86
 Modal emotions, 54
 Mood
 - emotion and, 2
- Motivation
 - emotion and, 80
 - empathy and, 23
- Motivational User Interface (MUI), 17
 Multimedia quality
 - evaluation of, 16
 - objective measure, 11
 - stress and, 16
 - subjective measure, 10
- Myers-Briggs Type Indicator, 199
- Nass, C., 173, 196, 198, 202, 223, 225
 Natural language generation
 - with affect, 207–217
- NIMIS Project (Networked Interactive Media In Schools), 21, 65
 - NIMIS Classroom, 22
- Nonverbal signals
 - meaning of, 191
- O'Malley, C., 15
 OCC model, 4, 5, 65, 70, 151–152, 155, 173
 Opportunistic behaviour, 90
 Ortony, A., 65, 173
- Pacman* game, 176
 Paiva, Ana, 1, 64
 Pelachaud, Catherine, 6, 182
 Perception
 - focus of attention and, 125
- Peripheral nervous system (PNS), 11
 Perlin, K., 198
 Personality, 105–106
 - and the Five Factor Model (FFM), 151–152
 - behaviour and, 196–202
 - believability, 173
 - *a definition*, 151
 - dimensions, 198
 - emotion and, 105–106, 154–155, 174
 - gestures and, 201
 - in animated lifelike characters, 150
 - in the Inhabited Market Place, 157–158
 - in the Presence project, 160–162
 - in the Puppet project, 155
 - modeling, 173, 198–201
 - goal-based approach, 174
 - posture and, 201
 - traits, 105, 197
- Petz, 198
 PF.Magic, 198
 Picard, R., 89, 204
 Picard, Rosalind, 2–4, 219
 Poggi, Isabella, 6, 182
 Positive affirmation
 - empathy and, 23
- Posture
 - personality and, 201
- Primary appraisal, 50
 Puppet project, 37
 - affect in, 152

- avatar in, 153
- synthetic characters in, 152
- QUASS (QQuality ASsessment Slider), 10, 15
- Raising eyebrows, 186
- Rational agent
 - decision making and, 122
- Reactive behaviour, 90
- Reape, M., 208
- Reappraisal, 50
- Reeves, B., 196, 202, 223, 225
- Reilly, W., 176
- Reverence, 3, 220
- Rhetorical Structure Theory, 193
- Rist, Thomas, 150
- Rizzo, Paola, 6, 166, 223
- RoboCup, 144
- Rogers, C., 22, 26
- Sadness, 38
- Sartre, J.P., 139
- Sasse, Angela, 2, 3, 9
- Scheller, M., 139
- Scherer, K., 50, 54
- Schmidt, S., 54
- Secondary appraisal, 50
- Self
 - sense of, 26
- Sengers, P., 177
- Sentiment
 - emotion and, 2
- Sexual drive, 90
- Shame, 80–88
 - beliefs and goals for, 82–83
- Simmel, M., 172
- Simon, H., 122, 146
- Situated Agent, 90
- Sloman, Aaron, 116
- Smile
 - meanings of, 51
- Solomon, R., 139
- Spinoza, 98
- Stork, D., 226
- Stress, 2
 - blood volume pulse and, 12
 - galvanic skin resistance and, 12
 - heart rate and, 12
 - measurement of, 11–13
 - user cost, 11
- Surprise, 38, 186
- Swagerman, J., 88
- Synthetic characters, *see* Animated lifelike characters
- Synthetic Emotions, 224
- Synthetic faces, 183
- Synthetic facial displays, 49
- TAGUS, 68
- Talking faces, 182
- Teatrix*, 4
 - and affective user modeling, 65
- Tetris, 224
- The Inhabited Market Place, 156
- The Presence Project, 158–162
- Thompson, S., 193
- Transactional, 50
- Trust, 77, 80–88, 225
 - attribution of, 84
 - Cognitive anatomy of, 83
 - degrees of, 85
 - implicit, 87
 - implicit and affective forms of, 85
 - lack of, 86
- UM-toolkit, 68
- User cost, 16–17
- User interfaces, 166
- User model
 - contents of, 68
 - emotional profile of user, 68
 - user attitudes and standards, 71
 - user emotions, 68
 - for affective text generation, 214–216
- Values, 5, 130
 - Ontology of, 130
- Velasquez, J., 118
- Virtual actors, 167
- Virtual Puppet Theater Architecture, 154
- Voice tone
 - empathy and, 23
- Walker, M., 208, 211
- Wang, L., 38
- Wehrle, Thomas, 3, 49, 54, 222
- Weizenbaum, J., 172
- Wiggins, 198
- Wilson, Gillian, 2, 3, 9, 10
- Wood, D., 25, 67

Author Index

- Allen, Steve 150
André, Elisabeth 150
Antunes, Luis 121
- Ball, Gene 196
Bogner, Myles 107
Bozinovski, Stevo 138
Breese, Jack 196
Brna, Paul 21
- Castelfranchi, Cristiano 76
Coelho, Helder 121
Cooper, Bridget 21
- Franklin, Stan 107
- Gebhard, Patrick 150
George, Pat 35
Grasso, Floriana 204
- Kaiser, Susanne 49
Klesen, Martin 150
- Machado, Isabel 64
Martinho, Carlos 64
Martins, Alex 21
McCauley, Lee 107
McIllhagga, Malcolm 35
- Paiva, Ana 1, 64
Pelachaud, Catherine 182
Picard, Rosalind 219
Poggi, Isabella 182
- Rist, Thomas 150
Rizzo, Paola 166
Rosis, Fiorella de 204
- Sasse, M. Angela 9
- Wehrle, Thomas 49
Wilson, Gillian M. 9