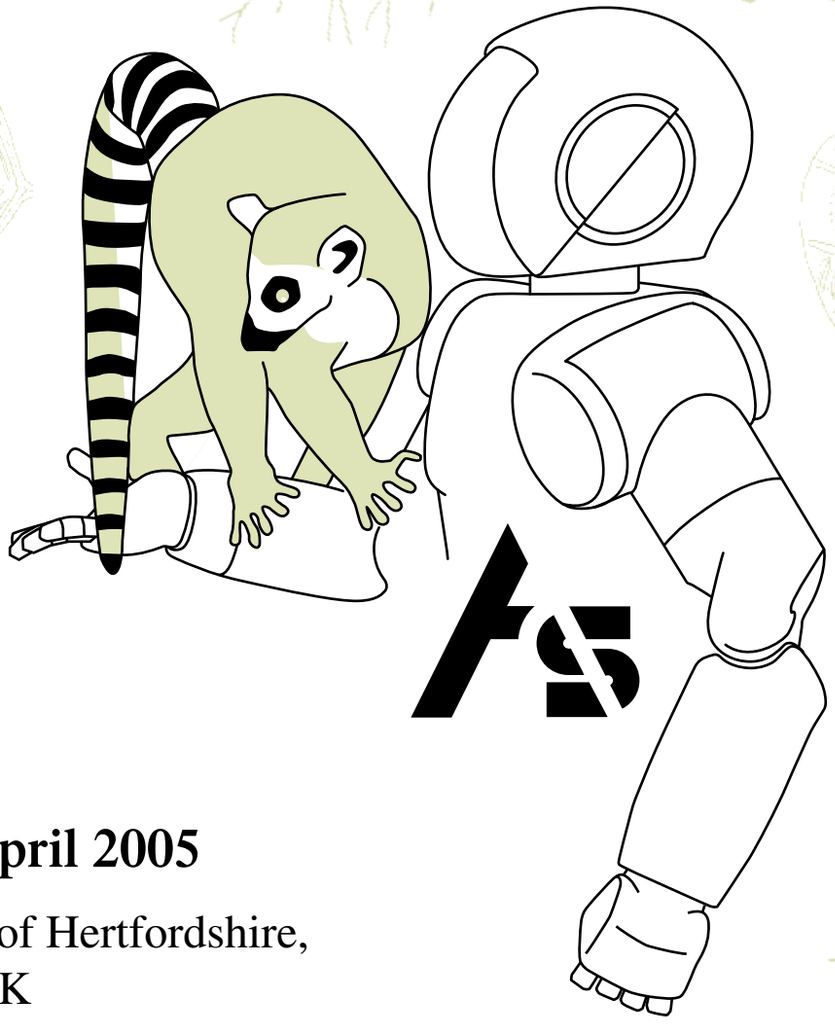
AISB'05: Social Intelligence and Interaction
in Animals, Robots and Agents

Proceedings of the Symposium on Agents that Want and Like: Motivational and Emotional Roots of Cognition and Action



12 - 15 April 2005

University of Hertfordshire,
Hatfield, UK

SSAISB 2005 Convention

AISB UH EPSRC
Engineering and Physical Sciences
Research Council

AISB'05 Convention

Social Intelligence and Interaction in Animals, Robots and Agents

12-15 April 2005
University of Hertfordshire, Hatfield, UK

Proceedings of the Symposium on

Agents that Want and Like:

Motivational and Emotional Roots of
Cognition and Action

Published by

The Society for the Study of Artificial Intelligence and the Simulation of Behaviour
www.aisb.org.uk

Printed by

The University of Hertfordshire, Hatfield, AL10 9AB UK
www.herts.ac.uk

Cover Design by Sue Attwood

ISBN 1 902956 41 7

AISB'05 Hosted by

The Adaptive Systems Research Group
adapsys.feis.herts.ac.uk

The AISB'05 Convention is partially supported by:

The proceedings of the ten symposia in the AISB'05 Convention are available from SSAISB:

Second International Symposium on the Emergence and Evolution of Linguistic Communication (EELC'05)
1 902956 40 9

Agents that Want and Like: Motivational and Emotional Roots of Cognition and Action
1 902956 41 7

Third International Symposium on Imitation in Animals and Artifacts
1 902956 42 5

Robotics, Mechatronics and Animatronics in the Creative and Entertainment Industries and Arts
1 902956 43 3

Robot Companions: Hard Problems and Open Challenges in Robot-Human Interaction
1 902956 44 1

Conversational Informatics for Supporting Social Intelligence and Interaction - Situational and Environmental Information Enforcing Involvement in Conversation
1 902956 45 X

Next Generation approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment
1 902956 46 8

Normative Multi-Agent Systems
1 902956 47 6

Socially Inspired Computing Joint Symposium (Memetic theory in artificial systems & societies, Emerging Artificial Societies, and Engineering with Social Metaphors)
1 902956 48 4

Virtual Social Agents Joint Symposium (Social presence cues for virtual humanoids, Empathic Interaction with Synthetic Characters, Mind-minding Agents)
1 902956 49 2

Table of Contents

The AISB'05 Convention
Social Intelligence and Interaction in Animals, Robots and Agents

Above all, the human animal is social. For an artificially intelligent system, how could it be otherwise?

We stated in our Call for Participation "The AISB'05 convention with the theme *Social Intelligence and Interaction in Animals, Robots and Agents* aims to facilitate the synthesis of new ideas, encourage new insights as well as novel applications, mediate new collaborations, and provide a context for lively and stimulating discussions in this exciting, truly interdisciplinary, and quickly growing research area that touches upon many deep issues regarding the nature of intelligence in human and other animals, and its potential application to robots and other artefacts".

Why is the theme of Social Intelligence and Interaction interesting to an Artificial Intelligence and Robotics community? We know that intelligence in humans and other animals has many facets and is expressed in a variety of ways in how the individual in its lifetime - or a population on an evolutionary timescale - deals with, adapts to, and co-evolves with the environment. Traditionally, social or emotional intelligence have been considered different from a more problem-solving, often called "rational", oriented view of human intelligence. However, more and more evidence from a variety of different research fields highlights the important role of social, emotional intelligence and interaction across all facets of intelligence in humans.

The Convention theme *Social Intelligence and Interaction in Animals, Robots and Agents* reflects a current trend towards increasingly interdisciplinary approaches that are pushing the boundaries of traditional science and are necessary in order to answer deep questions regarding the social nature of intelligence in humans and other animals, as well as to address the challenge of synthesizing computational agents or robotic artifacts that show aspects of biological social intelligence. Exciting new developments are emerging from collaborations among computer scientists, roboticists, psychologists, sociologists, cognitive scientists, primatologists, ethologists and researchers from other disciplines, e.g. leading to increasingly sophisticated simulation models of socially intelligent agents, or to a new generation of robots that are able to learn from and socially interact with each other or with people. Such interdisciplinary work advances our understanding of social intelligence in nature, and leads to new theories, models, architectures and designs in the domain of Artificial Intelligence and other sciences of the artificial.

New advancements in computer and robotic technology facilitate the emergence of multi-modal "natural" interfaces between computers or robots and people, including embodied conversational agents or robotic pets/assistants/companions that we are increasingly sharing our home and work space with. People tend to create certain relationships with such socially intelligent artifacts, and are even willing to accept them as helpers in healthcare, therapy or rehabilitation. Thus, socially intelligent artifacts are becoming part of our lives, including many desirable as well as possibly undesirable effects, and Artificial Intelligence and Cognitive Science research can play an important role in addressing many of the huge scientific challenges involved. Keeping an open mind towards other disciplines, embracing work from a variety of disciplines studying humans as well as non-human animals, might help us to create artifacts that might not only do their job, but that do their job right.

Thus, the convention hopes to provide a home for state-of-the-art research as well as a discussion forum for innovative ideas and approaches, pushing the frontiers of what is possible and/or desirable in this exciting, growing area.

The feedback to the initial Call for Symposia Proposals was overwhelming. Ten symposia were accepted (ranging from one-day to three-day events), organized by UK, European as well as international experts in the field of Social Intelligence and Interaction.

- Second International Symposium on the Emergence and Evolution of Linguistic Communication (EELC'05)
- Agents that Want and Like: Motivational and Emotional Roots of Cognition and Action
- Third International Symposium on Imitation in Animals and Artifacts
- Robotics, Mechatronics and Animatronics in the Creative and Entertainment Industries and Arts
- Robot Companions: Hard Problems and Open Challenges in Robot-Human Interaction
- Conversational Informatics for Supporting Social Intelligence and Interaction - Situational and Environmental Information Enforcing Involvement in Conversation
- Next Generation Approaches to Machine Consciousness: Imagination, Development, Intersubjectivity, and Embodiment
- Normative Multi-Agent Systems
- Socially Inspired Computing Joint Symposium (consisting of three themes: Memetic Theory in Artificial Systems & Societies, Emerging Artificial Societies, and Engineering with Social Metaphors)
- Virtual Social Agents Joint Symposium (consisting of three themes: Social Presence Cues for Virtual Humanoids, Empathic Interaction with Synthetic Characters, Mind-minding Agents)

I would like to thank the symposium organizers for their efforts in helping to put together an excellent scientific programme.

In order to complement the programme, five speakers known for pioneering work relevant to the convention theme accepted invitations to present plenary lectures at the convention: Prof. Nigel Gilbert (University of Surrey, UK), Prof. Hiroshi Ishiguro (Osaka University, Japan), Dr. Alison Jolly (University of Sussex, UK), Prof. Luc Steels (VUB, Belgium and Sony, France), and Prof. Jacqueline Nadel (National Centre of Scientific Research, France).

A number of people and groups helped to make this convention possible. First, I would like to thank SSAISB for the opportunity to host the convention under the special theme of *Social Intelligence and Interaction in Animals, Robots and Agents*. The AISB'05 convention is supported in part by a UK EPSRC grant to Prof. Kerstin Dautenhahn and Prof. C. L. Nehaniv. Further support was provided by Prof. Jill Hewitt and the School of Computer Science, as well as the Adaptive Systems Research Group at University of Hertfordshire. I would like to thank the Convention's Vice Chair Prof. Chrystopher L. Nehaniv for his invaluable continuous support during the planning and organization of the convention. Many thanks to the local organizing committee including Dr. René te Boekhorst, Dr. Lola Cañamero and Dr. Daniel Polani. I would like to single out two people who took over major roles in the local organization: Firstly, Johanna Hunt, Research Assistant in the School of Computer Science, who efficiently dealt primarily with the registration process, the AISB'05 website, and the coordination of ten proceedings. The number of convention registrants as well as different symposia by far exceeded our expectations and made this a major effort. Secondly, Bob Guscott, Research Administrator in the Adaptive Systems Research Group, competently and with great enthusiasm dealt with arrangements ranging from room bookings, catering, the organization of the banquet, and many other important elements in the convention. Thanks to Sue Attwood for the beautiful frontcover design. Also, a number of student helpers supported the convention. A great team made this convention possible!

I wish all participants of the AISB’05 convention an enjoyable and very productive time. On returning home, I hope you will take with you some new ideas or inspirations regarding our common goal of understanding social intelligence, and synthesizing artificially intelligent robots and agents. Progress in the field depends on scientific exchange, dialogue and critical evaluations by our peers and the research community, including senior members as well as students who bring in fresh viewpoints. For social animals such as humans, the construction of scientific knowledge can't be otherwise.

Beppu, Japan.

Dedication:

I am very confident that the future will bring us increasingly many instances of socially intelligent agents. I am similarly confident that we will see more and more socially intelligent robots sharing our lives. However, I would like to dedicate this convention to those people who fight for the survival of socially intelligent animals and their fellow creatures. What would 'life as it could be' be without 'life as we know it'?

Kerstin Dautenhahn

Professor of Artificial Intelligence,
General Chair, AISB'05 Convention *Social Intelligence and Interaction in Animals, Robots and Agents*

University of Hertfordshire
College Lane
Hatfield, Herts, AL10 9AB
United Kingdom

Symposium Preface
Agents that Want and Like:
Motivational and Emotional Roots of Cognition and Action

SYMPOSIUM OVERVIEW

The recent upsurge of interest in emotion in artificial intelligence has given rise to a number of workshops organized over the last years at international and national conferences in Europe (including several editions of the AISB Convention), USA and Japan. These meetings have tended to focus on topics such as the effects of emotion on cognitive functions (following the "paradigm shift" much stressed in recent years that considers emotion as a necessary element of intelligence and inseparable from cognition), emotion-based architectures, and emotions and their expression in social interaction. Other related affective phenomena such as personality and moods have also been the object of workshops, in particular within the human-computer interaction and user modeling communities.

At the same time motivation, even though overlooked for some time under the influence of behaviorism, is a topic of much interest in areas such as adaptive behavior and action selection, although social motivation remains for the most part unexplored in AI.

The interplay between emotion and motivation and their roles as "driving forces" underlying cognition and action (and particularly social intelligence and interaction) has however been largely neglected so far. Although "motivation and emotion" often constitutes one of the topics of interest at numerous AI conferences, researchers, with very few exceptions, tend to address only one of these topics and, to our knowledge, no workshop or symposium has explicitly focused on this important issue of their interrelationships and how they affect the way in which agents perceive, conceptualize and relate to the world and other agents. This symposium aims to redress this imbalance and to raise awareness of the importance of this topic among researchers interested in affect modeling and more generally in affect.

Motivation and emotion are indeed highly intertwined phenomena (e.g., emotions are often very powerful motivational factors; motivation can be seen as a consequence of emotion and viceversa, etc.) and it is not always easy to establish clear boundaries between them. Both types of phenomena are grouped under the broader category of "affect", traditionally distinguished from "cold" cognition. They lie at the heart of autonomy, adaptation, and social interaction in both biological and artificial agents. They also have a powerful and wide-ranging influence on many aspects of cognition and action. However, their roles are often considered to be complementary – as a first approximation, motivation would be concerned with the internal and external factors involved in the establishment of "goals" and the initiation and execution of goal-oriented action, whereas emotion is rather concerned, among other critical factors, with evaluative aspects of the relation between an agent and its environment.

This symposium proposes to investigate the roles and mutual interactions of motivation and emotion in influencing different aspects of cognition and action in biological and artificial agents that interact with their physical and social environment, as part of the 2005 AISB Convention general theme "Social Intelligence and Interaction in Animals, Robots and Agents." The nature of this topic necessitates a highly multi-disciplinary symposium; consequently, we have invited contributions from different relevant disciplines such as psychology, biology, neuroscience, ethology, sociology and philosophy, in addition to AI and robotics.

Further information about the symposium and the 2005 AISB Convention can be found at http://homepages.feis.herts.ac.uk/~comqlc/emotivation_aisb05 and at http://aisb2005.feis.herts.ac.uk.

This symposium would not have been possible without the collaboration of the members of the Program Committee: Orlando Avila-García (University of Hertfordshire, UK), Ruth Aylett (Heriot-Watt University, UK), Cynthia Breazeal (MIT, USA), Joanna Bryson (University of Bath, UK), Dylan Evans (University of the West of England, UK), Philippe Gaussier (University of Cergy-Pontoise, France), Steve Grand (Cyberlife Research Ltd., UK), Chris Melhuish (University of the West of England, UK), Jean-Arcady Meyer (LIP6, France), Jacqueline Nadel (CNRS & Hôpital de la Salpetrière, France), Paolo Petta (ÖFAI & Medical University of Vienna, Austria), Tony Prescott (University of Sheffield, UK), and David Sander (University of Geneva, CH). Thanks to all of them for their participation in the review process and their suggestions, feedback and various contributions regarding the symposium topic and organization. Thanks also to the Society for the Study of Artificial Intelligence and Simulation of Behaviour (SSAISB) and to the AISB'05 Convention team for providing an excellent framework for this symposium and for their support and help. Robert Marsh at the University of Hertfordshire provided additional help with the symposium proceedings. Particular thanks to Dylan Evans, initially symposium co-chair, who withdrew from his active role in the organization of the symposium due to unforeseen circumstances: his work during all these months is sincerely acknowledged.

Lola Cañamero
Symposium Chair

Emotion as a Motivated Behavior

George Ainslie

Veterans Affairs Medical Center, Coatesville and Temple Medical School
116A VA Medical Center, Coatesville, PA 19320
Email Ainslie@Coatesville.va.gov

Abstract

Emotions are conventionally treated as automatic processes that flow reflexively from assessments of reality. The assumption that future reward is discounted in standard percent-per-unit-time (exponential) discount curves has prevented recognition that emotions are at most quasi-automatic, and might be reward-dependent even when subjectively involuntary Substantial evidence that the basic discount curve is not exponential but hyperbolic makes possible a model in which even involuntary, negative emotions compete in a single internal marketplace of reward. A crude mechanical illustration of this model is described.

1 Introduction

Emotions are widely recognized to be motivating (Ortony *et.al*., 1988)—The name comes from the same Latin root, *movere,* the verb to move, the past participle of which is *motus*. However, they are thought of as being themselves unmotivated, rather as being imposed by the same process of classical conditioning to which most involuntary behaviors are attributed.

Certainly the major emotions have invariant features, are known to have specific brain circuits using specific neurotransmitters (Panksepp, 2000), and can even be induced by electrical brain stimulation (Delgado, 1969). In the original behaviorist model of emotion it was evoked as a conditioned response to innately determined stimuli (Watson, 1924). However, it proved to be hard to trace the emotional impact of a stimulus to a conditioning event. Even in the laboratory fear is the only emotion that has been conditioned; actual phobias are rarely a consequence of trauma involving the object feared, and trauma rarely leads to phobia (Rachman, 1977). The belief that an emotion is determined by a distant releasing stimulus linked to the immediate occasion by a chain of associations was a reasonable guess, but with little evidence behind it.

Later ideas of what induces emotion have been less specific, but still imply that it is driven by external givens that a person encounters—if not innately determined releasing stimuli, then belief that she faces a condition that contains these stimuli. Emotion is still a reflex of sorts, albeit usually a cognitively triggered reflex, a passive response to events outside of her control—hence "passion" as opposed to "action." In reviewing current cognitive theory, Frijda notes that the trigger may be as nonspecific as "whether and how the subject has appraised the relevance of events to concerns, and how he or she has appraised the eliciting contingency (2000, p. 68);" but this and the other theories of induction that he covers still involve an automatic response to the motivational consequences of the event, not a choice based on the motivational consequences of the emotion itself. Even though emotions all have such consequences, "the individual does not produce feelings of pleasure or pain at will, except by submitting to selected stimulus events (*ibid.,* p. 63)." That is, all emotions reward or punish, but they are said not to be chosen because of this consequence. In every current theory they are not chosen at all, but evoked.

2 Emotions can be shaped…

The widespread agreement that emotions are automatic ignores both common experience and a fair amount of data. Granted that emotions are usually *occasioned* by events outside of your voluntary control; the theory that they are *governed* by such events runs afoul of the widespread acknowledgment that they are trainable: You can "swallow" your anger or

"nurse" it, learn to inhibit your phobic anxiety (Marks & Tobena, 1990) or panic (Clum *et.al*., 1993; Kilic *et.al*., 1997) instead of "surrendering to it," limit your grief (Ramsay, 1997) instead of "wallowing in it," refrain from rejoicing or "give yourself over to it." Techniques to foster or inhibit emotions in everyday life have been described (Parrott, 1991), as has their use in preparing yourself for particular tasks (Parrott, 1993). Many schools of acting teach an ability to summon emotion deliberately (e.g. McGaw, 1966; Strasberg, 1988), because even in actors actual emotion is more convincing than feigned emotion (Gosselin *et.al*., 1998). The frequent philosophical assertion that emotions have a moral quality—good or bad (e.g. Hume as presented by Baier, 1991)—implies motivated participation; some philosophers have gone so far as to call the passions voluntary (e.g. Sartre, 1939/1948). In sum, emotions show signs of being goal-directed processes that are ultimately selected by their consequences, not their antecedents. That is, they are at least partially in the realm of reward-governed behaviors, not conditioned responses; they are pulled by incentives rather than pushed by stimuli. Even "negative" emotions like fear and grief seem to be urges that lure you into participating in them, rather than automatically imposed states. Conversely, the fact that emotions are usually involuntary does not mean that they are not selected by reward; after all, reward can even shape behavior during sleep (Granda & Hammack, 1961).

Examples of producing emotions deliberately are usually dismissed as examples of self-conditioning. Actors, for instance, use rehearsal of significant emotional memories to learn the necessary control, and psychotherapists often use guided imagery to influence emotions. According to conditioning theories you find the right conditioned stimulus and provoke your own reflex with it, like hitting your own knee with a rubber hammer to produce a jerk. It is true that in a given instance the goal-directed, or *operant,*[1] sequence of

cue→ response→ reward

can always be interpreted as the classically conditioned sequence of

conditioned stimulus→ conditioned response→ unconditioned or lower-order conditioned stimulus

and vice versa. However, if the conditioning stimulus is not repeated on successive trials, a true conditioned response will extinguish.[2] The memory or image will stop evoking the emotion. If the response grows and comes more readily, like the actor's emotion as she learns to summon it, it must have come under the control of a different selection agent, which probably means that it has been learned as an operant behavior. Learning to induce an emotion follows the same course as a bulimic's learning to vomit at will—the gagging stimulus of a spoon or finger becomes less and less necessary, and eventually can be dispensed with altogether.

2.1 …but how if by reward?

However, theoretical problems implicit in the concept of reward have been an obstacle to building an operant model of emotion. These theoretical problems follow from the conventional utility-based model of motivation. If you could produce "feelings of pleasure or pain at will," why not overdose on the pleasure and skip the pain, without regard to the outside world? If an emotion is aversive and avoidable, what induces people to entertain it? If an emotion is pleasurable and readily accessible, what keeps people from indulging in it *ad lib*?

3 Hyperbolic discounting supplies a mechanism

A solution has been unavailable because of a universal but almost certainly false assumption about how we evaluate future incentives. It is now well documented that both people and nonhuman animals have a robust tendency to devalue expected incentives in a hyperbolic curve. Such a curve represents a radical departure from the exponential curve that has been the explicit assumption of behavioral psychology and classical economics, and is implied by the "rational choice theory" that has become the norm in all behavioral sciences that depend on utility theory (Sugden, 1991; Cooter & Ulen, 2000).

[1] "Operant" is the favored term in behavioral psychology for "governed by differential reward and/or punishment."

[2] I have argued elsewhere that all "conditioned" responses can be understood as operant instead (1992, pp. 39-48; 2001, pp. 19-22), but I am not assuming that here. "Conditioned appetite" as a mechanism of preference reversal is analyzed in Ainslie, 2005.

3.1 Evidence that discounting is hyperbolic

Four kinds of experiment have demonstrated this phenomenon:

3.1.1 Goodness of fit

Given choices between rewards of varying sizes at varying delays, both human and nonhuman subjects express preferences that fit curves of the form,

$$V = A / (1 + kD)$$

a hyperbola, better than the form,

$$V = A\, e^{kD}$$

an exponential curve (where V is motivational value, A is amount of reward, D is delay of reward from the moment of choice, and k is a constant expressing impatience; Green, Fry & Myerson, 1994;Grace, 1996; Kirby, 1997; Mazur 2001). It has also been observed that the incentive value of small series of rewards is the sum of hyperbolic discount curves from those rewards (Mazur, 1986; Brunner & Gibbon, 1995).

3.1.2 Preference reversal

Given choices between smaller-sooner (SS) rewards and larger-later (LL) ones available at a constant lag after the SS ones, subjects prefer the LL reward when the delay before both rewards is long, but switch to the SS reward as it becomes imminent, a pattern that would not be seen if the discount curves were exponential (Ainslie & Herrnstein, 1981; Green *et.al*., 1981;Ainslie & Haendel, 1983; Kirby & Herrnstein, 1995). Where anticipatory dread is not a factor (with nonhumans or with minor pains in humans), subjects switch from choosing SS aversive stimuli to LL ones as the SS ones draw near (Solnick *et.al.*, 1980; Novarick, 1982; Dinsmoor, 1998).

3.1.3 Precommitment

Given choices between SS rewards and LL ones, nonhuman subjects will sometimes choose an option available in advance that prevents the SS alternative from becoming available (Ainslie, 1974; Hayes *et.al*., 1981). The converse is true of punishments (Deluty *et.al*., 1983). This design has not been run with human subjects, but it has been argued that illiquid savings plans and other choice-reducing devices serve this purpose (Laibson, 1997). Such a pattern is predicted by hyperbolic discount curves, while conventional utility theory holds that a subject has no incentive to reduce her future range of choices (Becker & Murphy, 1988).

3.1.4 Stabilization by bundling

When a whole series of LL rewards and SS alternatives must be chosen all at once, both human (Kirby & Guastello, 2001) and nonhuman (Ainslie & Monterosso, 2003a) subjects choose the LL rewards more than when each SS vs. LL choice can be made individually. The effect of such *bundling* of choices is predicted by hyperbolic but not exponential curves.

4 Overvaluation of immediate reward structures the emotions

The hyperbolic shape of the discount curve from delayed rewards makes possible an answer to the question raised above: What would make organisms entertain painful experiences, or limit their indulgence in pleasurable ones?

4.1 ”Negative” emotions

The argument for how negative emotions could be motivated behaviors involves the commonalities of aversive emotions and addictive rewards (Ainslie, 2001, pp. 90-104). Although both are usually avoided from a distance, both are seductive when they might occur in the near future. That is, however much you know that a binge will cost more than it is worth or that a fear is unfounded, it is sometimes hard not to participate in them.

Addictive behaviors can be well explained by imminent highs that, because of hyperbolic discounting, are valued temporarily above the more delayed rewards of sobriety (Vuchinich & Simpson, 1998; Mitchell, 1999). How the opposite rewarding and unrewarding incentives for negative emotions are compounded to attract attention but deter approach in general is still unclear. The similarity to addictive behaviors suggests that the urge to succumb to panic, anger, anguish, and even physical pain might be based on a rapidly recurring but very brief reward, lasting long enough to command attention but not deliberate choice, and fused in perception with longer, unrewarding

consequences to form an experience both vivid and aversive (Ainslie, 1992, pp. 100-114). Thus people who often encounter fearful situations—or who have a low fear threshold—sometimes learn to resist the urge to panic (Clum *et.al.*, 1993), but find it hard to do so despite an awareness that if they do not, panic will quickly prove to be the more aversive response.

4.2 "Positive" emotions

Although emotions are physically available, something makes them less intense in proportion as the occasion for them is arbitrary. To the extent that someone learns to access them at will, doing so makes them pale, mere daydreams. Even an actor needs to focus on appropriate occasions to bring them out with force. But what properties must an event have in order to serve as an occasion for emotion? The fact that there's no physical barrier opposing free access to emotions raises the question of how emotional experiences become the objects of often arduous striving, goods that seem to be in limited supply. That is, how do you come to feel as if you have them passively, as implied by their synonym, "passions?"

With the positive emotions, the basic question is, how does your own behavior become scarce? I'll divide it into two parts: Why would you want a behavior of yours to become scarce, that is, to limit your free access to it? And given that this is your wish, how can you make it scarce without making it physically unavailable?

All kinds of reward depend on a readiness for it that's used up as reward occurs and that can't be deliberately renewed. This readiness is the potential for appetite. The properties of appetites are often such that rapid consumption brings an earlier peak of reward but reduces the total amount of reward that the appetite makes possible, so that we have an amount-vs.-delay problem. Where people-- or, presumably, any reward-governed organisms-- have free access to a reward that's more intense the faster it's consumed, they will tend to consume it faster than they should if they were going to get the most reward over time from that appetite. In a conflict of consumption patterns between the long and pleasant versus the brief but even slightly more intense, an organism that discounts the future hyperbolically is primed to choose brief but intense. Accordingly, emotional reward, indulged in *ad lib*, becomes unsatisfactory for that reason itself. To get the most out of any kind of reward, we must have-- or develop-- limited access to it.

With emotional rewards, the only way to stop your mind from rushing ahead is to avoid approaches that can be too well learned. Thus the most valuable occasions will be those that are either 1. uncertain to occur or 2. mysterious-- too complex or subtle to be fully anticipated, arguably the goal of art. To get the most out of emotional reward, you have to either gamble on uncertainty or find routes that are certain but that won't become too efficient. In short, your occasions have to stay surprising-- a property that has also been reported as necessary for activity in brain reward centers (e.g. Hollerman *et.al.*, 1998; Berns *et.al.*, 2001). Accordingly, surprise is sometimes said to be the basis of aesthetic value (Berlyne, 1974; Scitovsky, 1976). In modalities where you can mentally reward yourself, surprise is the only commodity that can be scarce.

People-- and presumably nonhuman animals-- wind up experiencing as emotion only those patterns that have escaped the habituation of voluntary access, by a selective process analogous to that described by Robert Frank for the social recognition of "authentic" emotions (1988): Expressions that are known to be intentionally controllable are disregarded, as with the false smile of the hypocrite. By this process of selection, emotion is left with its familiar guise as passion, something that has to come over you.

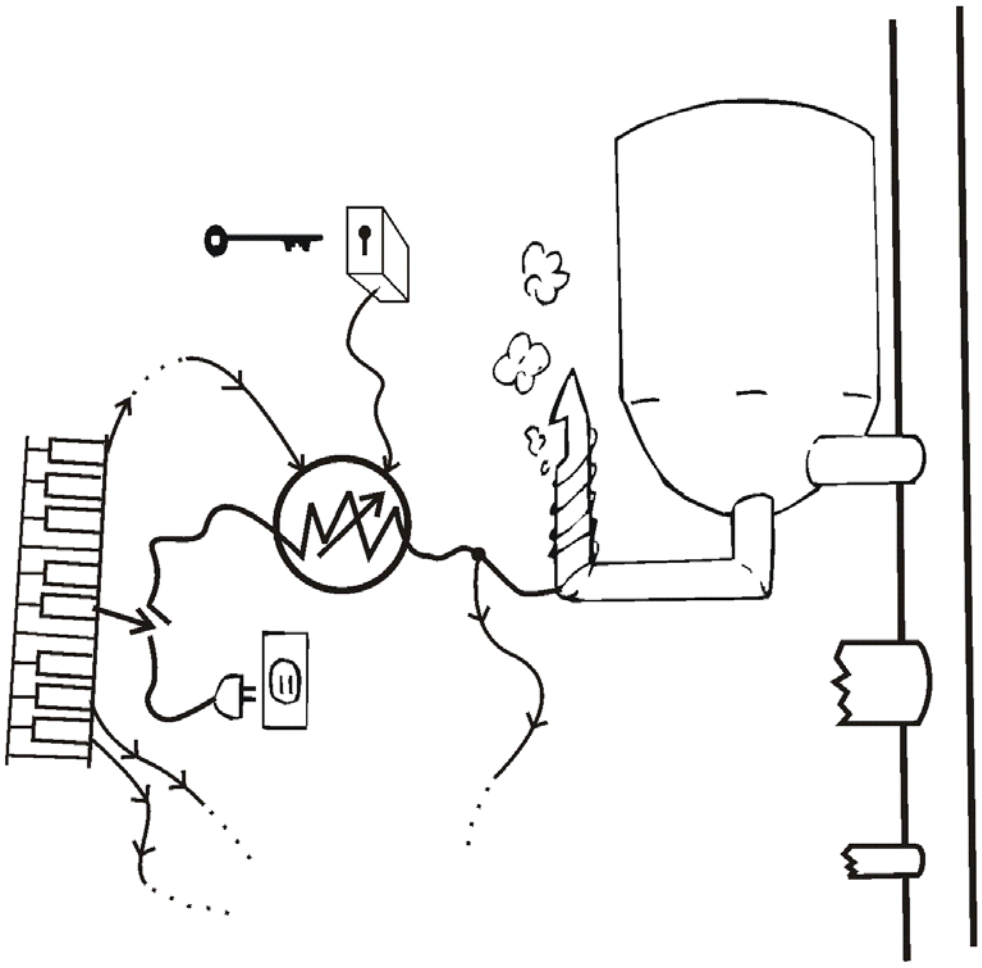
5 A motivational model of emotions

Hyperbolic discounting greatly simplifies the problem of modeling the emotions. With conventional, exponential curves, a person should be able to estimate what emotions will be most rewarding for what durations, and plan accordingly. To correct this picture to match the real world, a modeler has to impose negative emotions on the subject, and limit her access to positive emotions, by a combination of hardwired and conditioned reflexes. By contrast, hyperbolic discounting lets emotions be behaviors that compete in the common market of motivation. In such a model, emotions differ from deliberate but volatile behaviors like paying attention only in producing significant intrinsic reward. The patterns of this reward determine both emotions' quasi-involuntary property and

the motive to limit their occurrence—the negative emotions by an admixture of obligatory nonreward that overbalances their reward at all but very short distances, the positive emotions by the premature satiation that will occur unless the subject limits what occasions their occurrence.

5.1 The Demon at the Calliope

This situation can be portrayed by an automated model, and even a mechanical one. I will describe the latter for better illustration (cf. Ainslie, 1992, pp. 274-291). The individual is divided into a motivating part and a behaving part. The motivating part is the brain function that generates reward, modeled by the whistles of steam organ (circus calliope). The calliope has individual steam boilers heated by their own circuits—one for each separately satiable modality of emotion, such as anger, sexual arousal, laughter, and even grief and panic (figure). Other boilers exist for nonemotional options such as muscle movements. The behaving part is a demon who presses the calliope keys according to a strict instruction: "Choose the option that promises the greatest aggregate of loudness x duration, discounted hyperbolically to the present moment."



A single boiler heated by current that is controlled by one key of the calliope. The whistle can blow as long as it has heat and water; the water is replaced in the boiler at a rate determined by the diameter of the intake pipe. A rheostat governed by hardwired factors including turnkey stimuli and current flow in other boilers can modify current flow, and current flow can affect rheostats on other boilers. The loudness of the whistle is not a linear function of the amount of steam produced; it is disproportionately less at very low and very high values.

5.1.1 Properties of the calliope

Pressing a key sends electric current through heating coils around its boiler, causing release of steam through the whistle at a delay and over a time course determined by several factors:

- The shape of the boiler. Narrow necks limit loudness, and bigger tanks hold more water, modeling the potential intensity and duration of the emotion.
- The density of wiring around the boiler neck relative to its diameter. This models the speed of arousal.
- The amount of water in the boiler. This models physiological readiness for the emotion (something like "drive").
- The rate at which the demon presses the key. Pressing too slowly wastes the effort, too fast exceeds the whistle's sound-producing capacity and wastes steam.
- The diameter of intake pipe to the boiler, modeling the rate at which readiness regenerates
- The presence of turnkeys to the rheostat (variable resistor) in the heating wire, modeling the extent to which hardwired stimuli (e.g. pain) facilitate the emotion. Emotions vary in their readiness to occur without hardwired turnkey ("unconditioned") stimuli, and a given process varies among individuals, as in the traits of fear- or fantasy-proneness. This readiness is modeled by what is the lowest setting of the rheostat.
- Activity in the heating coils of other boilers that are hardwired to raise or lower this rheostat. For instance, pain might augment sexual arousal or decrease laughter.

5.1.2 The behavior of the model

The demon has whatever estimating ability the whole individual has, which I do not model further. Emotions are all wired for fast partial payoffs, although their long run payoffs are variable. Because of their fast payoffs they have a great ability to compete with other choices on the demon's keyboard. Because hyperbolic discounting makes curves from imminent payoffs disproportionately high, the demon will often be lured into negative emotions—those that do not have enduring payoffs and that lower the rheostat on other boilers—when a turnkey stimulus is present and/or readiness is high. For the same reason he will press wastefully and not get the most steam from the available water in positive emotions if he presses keys *ad lib.* Thus he will be motivated to tie his pressing to the appearance of adequately rare external cues.

5.2 The value of the model

A quantitatively accurate model would reflect the time course of neuronal processes, of course, most of which are still unknown. Even the sites of interaction of the components that I have illustrated are merely the simplest that will relate the dynamic of hyperbolic discounting to the known properties of drive and emotion. I do not pretend to fit the promising but still sketchy single neuron physiology and fMRI data that are beginning to emerge.

The point of this crude model is to add flesh to the bare mathematical fact that hyperbolic valuation curves describe the temporary dominance of some SS outcomes over some LL ones. That property makes possible a model that uses only one selective process (reward) instead of the conventional two (classical conditioning and reward), and that requires all learnable processes, even emotions, to compete in the single internal marketplace of motivation. A one-process model is not only more parsimonious than the conventional one, but also better fits the phenomenon of mixed emotions—the strangely addictive quality shown often by anger and sometimes even by grief and fear. Beyond that, as I have argued elsewhere (2001, pp. 175-186), a model of emotions that has stimuli serve as occasions for them rather than rather than control them makes possible dynamic theories of the psychological/social construction of facts and of empathy as a primary (not instrumental) good.

.

References

G. Ainslie. Impulse control in pigeons. *Journal of the Experimental Analysis of Behavior*, 21:485-489, 1974.

G. Ainslie. *Picoeconomics: The Strategic Interaction of Successive Motivational States within the Person.* Cambridge, 1992.

G. Ainslie. *Breakdown of Will*. Cambridge, 2001.

G. Ainslie. A selectionist view of the ego: Implications for self-control. In N. Sebanz and W. Prinz, Editors., *Disorders of Volition*. MIT, 2005.

G. Ainslie, and V. Haendel. The motives of the will. In E. Gottheil, K. Druley, T. Skodola, H. Waxman (Editors), *Etiology Aspects of Alcohol and Drug Abuse.* Charles C. Thomas, pp.119-140, 1983.

G. Ainslie, and R. Herrnstein. Preference reversal and delayed reinforcement. *Animal Learning and Behavior*, 9:476-482, 1981.

G. Ainslie, and J. Monterosso. Building blocks of self-control: Increased tolerance for delay with bundled rewards. *Journal of the Experimental Analysis of Behavior,* 79, 83-94, 2003.

A. Baier. *A Progress of Sentiments: Reflections on Hume's Treatise*. Harvard University, 1991.

G. Becker, and K. Murphy. A theory of rational addiction. *Journal of Political Economy*, 96:675-700, 1988.

D.E. Berlyne. *Studies in the New Experimental Aesthetics*. Hemisphere, 1974.

G.S. Berns, S.M. McClure, (??) et al. Predictability Modulates Human Brain Response to Reward, *Journal Of Neuroscience,* 21, 2793-2798, 2001.

D. Brunner, and J. Gibbon. Value of food aggregates: Parallel versus serial discounting. *Animal Behavior*, 50:1627-1634, 1995.

G.A. Clum, G.A. Clum, and R. Surls. A meta-analysis of treatments for panic disorder. *Journal of Consulting and Clinical Psychology*, 61:317-326, 1993.

R. Cooter, and T. Ulen. *Law and Economics*. Addison-Wesley, 2000.

J.M.R. Delgado. *Physical Control of the Mind: Toward a Psychocivilized Society*. Harper & Row, 1969.

M.Z. Deluty, W.G. Whitehouse, M. Mellitz, and P.N. Hineline. Self-control and commitment involving aversive events*. Behavior Analysis Letters*, 3:213-219, 1983.

J.A. Dinsmoor, J. A. Punishment. In W. T. O'Donohue (Editor), *Learning and Behavior Therapy*. Allyn & Bacon, 1998.

R.H. Frank. *Passions Within Reason*. Norton, 1988.

N.H. Frijda, The psychologists' point of view. In M. Lewis and J. M. Haviland-Jones (Editors), *Handbook of Emotions 2d Edition*. Guilford, 2000.

P. Gosselin, G. Kirouac, and F.Y. Dore. Components and recognition of facial expression in the communication of emotion by actors. *Journal of Personality and Social Psychology*, 68:83-96, 1998.

R. Grace. Choice between fixed and variable delays to reinforcement in the adjusting-delay procedure and concurrent chains. *Journal of Experimental Psychology: Animal Processes*, 22:362-383, 1996.

A. M. Granda and J. T. Hammack. Operant behavior during sleep. *Science* 133, 1485-1486, 1961.

L. Green, E.B. Fisher, Jr., S. Perlow, and L. Sherman. Preference reversal and self-control: Choice as a function of reward amount and delay. *Behaviour Analysis Letters,* 1, 43-51, 1981.

L. Green, A. Fry, and J. Myerson. Discounting of delayed rewards: A life-span comparison. *Psychological Science*, 5:33-36, 1994.

S.C. Hayes, J. Kapust, S.R. Leonard, and I. Rosenfarb. Escape from freedom: Choosing not to choose in pigeons. *Journal of the Experimental Analysis of Behavior,* 36, 1-7, (1981).

J. R. Hollerman, L. Tremblay, and W. Schultz. Influence of reward expectation on behavior-related neuronal activity in primate striatum. Journal of Neurophysiology 80, 947-963, 1998.

C. Kilic, H. Noshirvani, M. Basoglu, and L. Marks. Agoraphobia and panic disorder: 3.5 years after alprazolam and/or exposure treatment. *Psychotherapy and Psychosomatics*, 66:175-178, 1997.

K.N. Kirby. Bidding on the future: Evidence against normative discounting of delayed rewards. *Journal of Experimental Psychology: General*, 126:54-70, 1997.

K.N. Kirby, and B. Guastello. Making choices in anticipation of similar future choices can increase self-control. *Journal of Experimental Psychology: Applied*, 7:154-164, 2001.

K. N. Kirby and R.J. Herrnstein. Preference reversals due to myopic discounting of delayed reward. *Psychological Science*, 6:83-89, 1995.

D. Laibson. Golden eggs and hyperbolic discounting. *Quarterly Journal of Economics*, 62:443-479, 1997.

C. McGaw. *Acting is Believing: A Basic Method*. Holt, Rinehart, & Winston, 1966.

I. Marks, and A. Tobena. Learning and unlearning fear: A clinical and evolutionary perspective. *Neuroscience and Biobehavioral Reviews*, 14:365-384, 1990.

J.E. Mazur. Choice between single and multiple delayed reinforcers. *Journal of the Experimental Analysis of Behavior*, 46:67-77, 1986.

J.E. Mazur. Hyperbolic value addition and general models of animal choice. *Psychological Review*, 108:96-112, 2001.

S. Mitchell. Measures of impulsivity in cigarette smokers and nonsmokers. *Psychopharmacology*, 146:455-464, 1999.

D.J. Novarick. Negative reinforcement and choice in humans. *Learning and Motivation*, 13:361-377, 1982.

A. Ortony, G. Clore, and A. Collins. *The Cognitive Structure of Emotions*. Cambridge University, 1988

J. Panksepp. Emotions as natural kinds within the mammalian brain. In M. Lewis & J. M. Haviland-Jones, eds. *Handbook of Emotions 2d Edition*. Guilford, 2000.

W.G. Parrott. Mood induction and instructions to sustain moods. A test of the subject compliance hypothesis of mood congruent memory. *Cognition and Emotion*, 3:41-52, 1991.

W.G. Parrott. Beyond hedonism: Motives for inhibiting good moods and for maintaining bad moods. In D. M. Wegner & F. W. Pennebaker (Editors), *Handbook of Mental Control*, 278-305. Prentice Hall, 1993.

R.W. Ramsay. Behavioural approaches to bereavement. In S. Rachman & H. J. Eysenck (Editors), *The Best of Behavior Research and Therapy*. Pergamon, 1997.

S.J. Rachman. The conditioning theory of fear acquisition: A critical examination. *Behaviour Research and Therapy*, 15:375-388, 1977.

J.P. Sartre. *The Emotions: Sketch of a Theory* (B. Frechtman, trans.) Philosophical Library, 1939/1948.

J. Solnick, C. Kannenberg, D. Eckerman, and M. Waller. An experimental analysis of impulsivity and impulse control in humans. *Learning and Motivation*, 2:61-77. Review, 217-225, 1980.

L. Strasberg. *A Dream of Passion: The Development of the Method*. Dutton, 1988.

R. Sugden. Rational choice: a survey of contributions from economics and philosophy. *Economic Journal*, 101:751-785, 1991.

R.E. Vuchinich, and C.A. Simpson. Hyperbolic temporal discounting in social drinkers and problem drinkers. *Experimental and Clinical Psychopharmacology,* 6:292-305, 1998.

J.B. Watson. *Behaviorism*. The Peoples Institute, 1924.

Hormonal Modulation of Perception in Motivation-Based Action Selection Architectures

Orlando Avila-García*
*Adaptive Systems Research Group
Department of Computer Science
University of Hertfordshire
College Lane, Hatfield, Herts AL10 9AB, UK
O.Avila-Garcia@herts.ac.uk

Lola Cañamero†
†Adaptive Systems Research Group
Department of Computer Science
University of Hertfordshire
College Lane, Hatfield, Herts AL10 9AB, UK
L.Canamero@herts.ac.uk

Abstract

The *animat* approach to artificial intelligence proposes biologically-inspired control mechanisms for autonomous robots. One of the related subproblems is action selection or "what to do next". Many action selection architectures have been proposed. Motivation-based architectures implement a combination between internal and external stimuli to choose the appropriate behavior. Recent studies have pointed out that a second order mechanism to control motivation-based architectures would improve dramatically their performance. Drawing on the notion of biological hormones we have modeled two of the functionalities ascribed to them in order to improve the adaptivity of motivation-based architectures. We have tested our "hormone-like" mechanisms in dynamic and unpredictable robotic scenarios. We analyze the results in terms of interesting behavioral phenomena that emerge from the interaction of these artificial hormones with the rest of architectural elements.

1 Introduction

Within the "behavior-based" (Brooks, 1986; Steels and Brooks, 1995) or *animat* approach (Wilson, 1985; Meyer, 1995) to AI, the ultimate goal of an autonomous agent is survival in a given dynamic, unpredictable and possibly threatening environment. Following inspiration from models in biology, neuroscience and cybernetics, animat's survival needs are commonly represented as internal essential variables. In order to remain "alive" the animat must maintain homeostasis, i.e., keep the level of those essential variables within certain ranges of viable values (Ashby, 1952; Aubin, 2000). Since different courses of action can be taken to maintain homeostasis, one of the related subproblems is action selection (Maes, 1995), i.e., making a decision as to what behavior to execute in order to guarantee survival in a given environment and situation.

Many action selection architectures have been proposed (see Tyrrell (1993) or Guillot and Meyer (1994) for an overview). Following the behavior-based approach to robotics, architectures started to be essentially reactive. Later on it became apparent that some internal stimuli—e.g., the level of the essential variables—-were necessary in order to keep those internal variables within their ranges (Arkin, 1992). Following inspiration from ethology (Timbergen, 1951; McFarland, 1999) and motivational systems (McFarland, 1974; Toates, 1986), different ways of combining internal and external factors started to appear, proposing integration of those factors at different levels of the action selection process (Maes, 1991; Tyrrell, 1993; Blumberg, 1994; Spier and McFarland, 1997). In motivation-based architectures (Cañamero, 1997), motivations constitute tendencies to maintain homeostasis as a consequence of internal and external factors.

In previous studies, we compared different motivation-based action selection architectures (Avila-García and Cañamero, 2002), and we suggested that the cyclic fashion in which motivations are satisfied greatly influences the performance of the agent (Avila-García et al., 2003). In this paper, we show that the same action selection architecture, appropriate in certain static environments, does not perform viable activity cycles in environments with added dynamic complexities. We propose "hormone-like" mechanisms to adapt action selection architectures to changing and dynamic environmental circumstances. Such mechanisms modulate the sensory input of motivation-based architectures in order to adapt its decisions to the changing environmental circumstances. Our

modulatory mechanisms are based on the functional properties of hormones (Levitan and Kaczmarek, 1997).

Section 2 describes the architectural elements of our action selection architectures. Section 3 shows how hormonal modulation of the perception of external stimuli (exteroceptors) adapt the action selection architecture to a competitive scenario. Section 4 shows hormonal modulation of the perception of one internal essential variable (interoceptor) to adapt the architecture to a dynamic prey-predator scenario. Section 5 draws some conclusions.

2 Action Selection Architectures

Our motivation-based architectures consist of two layers—motivational and behavioral—linked through a synthetic physiology, leading to a two-step computation of intensity. This computation is parallel within each layer, but motivational intensity must be computed prior to the calculation of behavioral intensity, since the latter depends on the former. The motivational layer is made of motivational states that set the goals of the system—the tendency to satisfy bodily (physiological) or internal needs. The behavioral layer implements different ways in which those bodily needs can be satisfied. This distinction between motivations and behaviors is essential when implementing more than one behavior satisfying the same motivation (Toates, 1986).

The Physiology consists of a number of survival-related, homeostatically controlled **essential variables**—abstractions representing the level of internal resources that the agent needs in order to survive. They must be kept within a range of values for the robot to stay "alive," thus defining a physiological space (Sibly and McFarland, 1974) or viability zone (Ashby, 1952; Meyer, 1995) within which survival (continued existence) is guaranteed, whereas transgression of these boundaries leads to "death."

Motivations are abstractions representing tendencies to behave in particular ways as a consequence of internal and external factors (Toates, 1986). Internal factors are mainly (but not only) bodily or physiological deficits or needs ($0 \leq d_i \leq 1$), also traditionally known as "drives," that set urges to action to maintain the state of the controlled physiological variables within the viability zone. External factors are environmental stimuli or incentive cues ($0 \leq c_i \leq 1$) that allow to execute (consummatory) behaviors and hence to satisfy bodily needs. In our implementation, each motivation performs homeostatic control of one physiological variable. We have used the equation

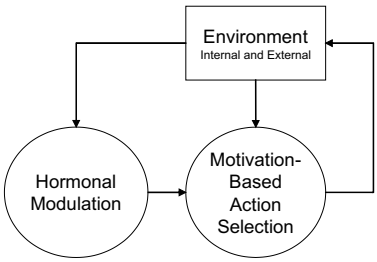


Figure 1: Model of hormonal modulation of a motivation-based action selection architecture.

proposed in (Avila-García et al., 2003) to combine cue and physiological deficit when computing motivations' intensities:

$$m_i = d_i + (d_i \times \alpha c_i) \quad (1)$$

In addition to physiological deficits (d_i) and the presence of environmental cues (c_i), we have introduced a weighting factor ($0 \leq \alpha \leq 1$) to change the relevance given to the external cue. Note that with $\alpha = 0$ the intensity of the cue does not have any effect on the motivation's intensity.

Behaviors are coarse-grained subsystems (embedding simpler actions) that implement different competencies, similarly to those proposed in (Maes, 1991; Cañamero, 1997). Following the classical distinction in ethology (McFarland, 1999) and more recent advances in neuroscience (Robbins and Everitt, 1999), motivated behaviors can be consummatory (goal-achieving and needing the presence of an incentive stimulus to be executed) or appetitive (goal-directed search for a particular incentive stimulus). In addition to modifying the external environment, the execution of a behavior has an impact on (increases or decreases) the level of specific physiological variables. Therefore they are a mechanism to maintain the state of the physiological variables within the viability zone.

2.1 Hormonal Control Layer

Our architectures contain a second order control layer in the form of hormone-like mechanisms (Figure 1). We take inspiration from models of hormonal control in neuroscience (Kravitz, 1988; Harris-Warrick et al., 1992) to modulate the sensory channels of motivation-based action selection architectures:

(a) Sensory inputs enhance the release of hormones that act at different levels of the nervous system: e.g. sensory elements.

Table 1: Motivations used. Physiological drive represents the tendency of the motivation to decrement the physiological deficit until the correspondent variable reaches a set point of 100.

Motivation	Drive	Limit	Set point	Ext. Stim.
m_{cold}	$\downarrow d_{temperature}$	0	100	c_{heat}
$m_{fatigue}$	$\downarrow d_{energy}$	0	100	c_{food}

Table 2: Behaviors used by the WTA architecture.

Behavior	Type	Stimulus	Effects on physiology
b_{avoid}	*Reflex.*	$s_{obstacle}$	+ 0.2 d_{temp}, + 0.2 d_{energy}
b_{warmup}	*Consum.*	c_{heat}	– 1.0 d_{temp}, + 0.3 d_{energy}
b_{feed}	*Consum.*	c_{food}	+ 0.3 d_{temp}, – 1.0 d_{energy}
b_{search}	*Appet.*	None	+ 0.2 d_{temp}, + 0.2 d_{energy}

(b) They act as gain-setting sensitization process that biases the output of the organism in particular directions.

(c) The organism now responds to particular sensory stimuli with an altered output appropriate to the new situation.

3 Modulation of Exteroceptors

In previous studies we analyzed different motivation-based action selection architectures within a static Two-Resource Problem (TRP) (Avila-García et al., 2003), where a single robot must maintain optimum *temperature* and *energy* levels consuming two resources available in the environment—*heat* and *food* respectively. Resources were static, i.e., their location did not change and they were always equally accessible. In that study we used a Lego Mindstorms robot, performing in a $1m \times 1m$ arena surrounded by a wall, with bright and black gradients representing heat and food sources respectively (see Figure 2). Tables 1 and 2 detail our particular implementation of the TRP. We used two environmental resources (incentive cues) that allow to satisfy two physiological needs, the deficits of which give rise to two motivational states. Motivations are satisfied by the execution of a consummatory behavior. The execution of each behavior reduces the deficit of one physiological variable, increasing the deficit of the other one.

Our extension of this problem, the Competitive Two-Resource Problem (CTRP) consists in the introduction of two robots in the same environment simultaneously performing their own TRP. The fact that they have to use the same resources to satisfy their needs introduces competition for those resources, as both robots might need access to the same resource at the same time. Thus, new forms of environmental complexity—availability and accessibility of resources—appear due to the interaction between robots. These forms of complexity effectively affect the stability and viability of activity cycles performed by the robots (Avila-García and Cañamero, 2004b).

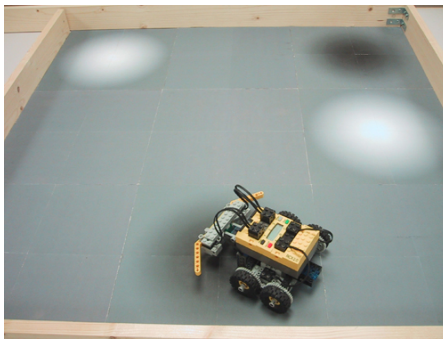


Figure 2: Our robotic Two-Resource Problem (TRP) scenario. The environment is a $1m \times 1m$ arena inhabited by one robot, two *heat* resources and two *food* resources. *Heat* and *food* resources are represented as brightness and darkness gradients on the floor of the arena.

Using activity cycles analysis (Avila-García et al., 2003) we can see that the action selection mechanism used within the TRP presents incoherences within the CTRP, performing pathological cycles of activities. Firstly, the robot can fall in a pathological sequence of opportunistic activities—consuming the same resource—that eventually can drive the robot to death (over-opportunism). Secondly, when one robot is located on top of a resource—i.e., consuming it—the other robot might bump into it and push it out of the resource. This will result in the interruption of the ongoing consummatory activity.

3.1 Hormonal Mechanism

Neurohormones interact with the nervous system to modulate behavioral output, acting at different levels and evoking a spectrum of different responses on different target neurons, in what has been called "neuromodulation" (Kravitz, 1988). In this section we are inspired by the ability of neuromodulators to extract different functionalities from the same anatomical neural circuit (Harris-Warrick et al., 1992; Levitan and Kaczmarek, 1997).

We propose a single "hormone-like" mechanism to solve the problems described above. Firstly, it re-

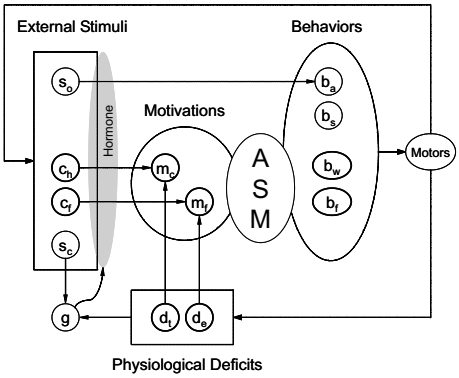


Figure 3: Proposed hormone-like modulatory mechanism for the action selection architecture. A gland (g) secretes the hormone, whose concentration (c_g) is a function of the stimulus *competitor* (s_c) and the Risk of Death (RoD) from physiological deficits.

duces opportunistic activities when there is any risk of death. By acting on the parameter α of equation 1—i.e., biasing the relevance given to external cues—the hormone reduces the perception of both incentive cues. Secondly, the hormone potentiates the competition skills of the robot by enhancing its capacity to push the other robot out the resources and not to be interrupted when consuming resources. This is obtained by cancelling the perception of obstacles ($s_{obstacle}$)—and hence the avoidance reflex behavior—when the robot is facing the competitor.

To achieve this twofold functionality, the concentration of hormone will be the function of the risk of death and the perception of the competitor. We define Risk of Death (RoD) as the inverse of the distance between physiological state (d_{temper}, d_{energy}) and lethal boundaries. The perception of the competitor is given by a new stimulus $0 \leq s_{competitor} \leq 1$. Hormone concentration will be computed as:

$$c_g = RoD + s_{competitor} \tag{2}$$

Finally, we have to define the relation between hormone concentration and the cancellation of the perception of incentive cues and obstacles. In order to achieve the first functionality, the cancellation of α is directly proportional to the increment in the hormone concentration. Therefore, when RoD rises α drops.

$$\alpha = min(1 - c_g, 0) \tag{3}$$

The second functionality is obtained by cancelling the perception of $s_{obstacle}$—i.e., bumpers—when the competitor is in front of the robot. There are two conditions that must be fulfilled to make a coherent pushing of the other robot. First, the robot must avoid getting engaged in fights when it has high RoD. Second, it must only bump blindly into the other robot, not against the walls of the arena. To produce that effect the cancellation of the bumpers must be at hormonal levels $c_g \simeq 1$ and $c_g \simeq 2$.

Note that the motivation-based action selection architecture has suffered no modification, apart from the fact that now one of its parameters (α, equation 1) at sensory level is modulated by the hormonal feedback mechanism.

3.2 Experiments

We have tested the robots in a total of 16 runs of 1200 steps each. Each step represents a loop of the action selection mechanism, taking $260ms$[1]. This means that each run lasts about 5 minutes.

The robot with hormone-like mechanism recovers the stability and viability of activity cycles. In this paper we focus on the interesting functionalities that emerge from the modulation of the perception of the robot's external stimuli; for a quantitative analysis in terms of viability indicators and activity cycles see (Avila-García and Cañamero, 2004b). The first functionality is stopping of consuming resources when the robot detects its competitor approaching. This could be interpreted by an external observer as abandonment of a situation (waiting for the other robot at the resource) that is disadvantageous to compete. Instead, the robot will leave the resource and go straightforward towards the competitor until it reaches it; at that moment, if there is some level of RoD, the bumpers of the robot will not be cancelled and it will avoid the competitor—showing a behavior that an observer could interpret as "fear" after evaluating the competitor. On the contrary, if there is not RoD, the hormonal system will cancel the bumpers and the robot will push the competitor unconditionally—as if it showed some sort of "aggression" against it. If we study the whole picture as external observers, the previous behavioral phenomena could well be interpreted as some sort of "protection of resources".

4 Modulation of Interoceptors

In this second part of the paper we extend the previous study by introducing a third physiological vari-

[1]Lego Mindstorms robots use a 16MHz microcontroller

Table 3: Extra motivation used in the H3RP.

Motivation	Drive	Limit	Set point	Ext. Stim.
m_{damage}	$\downarrow d_{integrity}$	0	100	c_{nest}

Table 4: Extra behavior used in the H3RP.

Behavior	Type	Stimulus	Effects on physiology
$b_{recover}$	*Consum.*	c_{nest}	+ 0.4 d_t, + 0.4 d_e, – 1.0 d_i

able, *integrity*, which is unpredictably reduced by the environment (see Table 3). We call our new test bed the Hazardous 3-Resource Problem (H3RP). This extra variable will work as a metaphor of the essential need any organism has to keep its tissue—the boundary between the organism and its environment—without damage. The environment reduces the robot's *integrity* level in a semi-unpredictable way through the action of an extra robot, the predator. The predator simply chases the prey robot and bumps into it, touching a wire ring that surrounds the prey's structure. The prey robot is able to recover the level of energy going to a specific place in the environment, the *nest* (see Table 4). Figure 4 shows both prey and predator within the arena.

In preliminary experiments we can observe that a motivation-based action selection mechanism does not perform well within the new framework (Avila-García and Cañamero, 2004a). Using activity cycles analysis we show that the behavior of the robot in the new environment is incoherent. The problem arises when the action selection mechanism pays low attention to the new motivation to *recover integrity* even when the predator is at sight. The probability to lose *integrity* rises when the predator is around, therefore the prey robot should predict that loss and anticipate the *recovering* of integrity.

A simple solution consists in using one of the existing sensors of the prey robot to detect the predator. This sensor must be the same one used to locate the nest. The problem of using that sensor is that it is fixed pointing forwards. Since the predator seldom passes in front of the prey, this extra stimulus ($s_{predator}$) will be too low to make any difference. However, our hormone makes the system more sensitive to *integrity* deficit after the detection of the predator. Hormonal secretion follows the detection of the stimulus $s_{predator}$ and increases the perceived *integrity* deficit. Using the hormone's temporal dynamics, the modulation will be acting in the system long time after the predator has disappeared from sensory inputs. The concentration of hormone thus modifies one of the inputs of the architecture and, as we will see in the next section, biases action selection. We thus make use of the temporal dynamics of an artificial hormone to maintain the stimulus $s_{predator}$ increasing the motivation to *recover* long after the predator is out of sight. Like in the previous case, we propose a "hormones-like" mechanism that acts again as a second order modulator of the motivation-based action selection architecture.

Figure 4: The prey robot *recovering* the level of *integrity*, facing the corner that represents the *nest*, about to be attacked by the predator.

4.1 Hormonal Mechanism

In this section we model another functionality ascribed also to hormones: long-term changes in behavior (Levitan and Kaczmarek, 1997). We have modeled hormonal temporal dynamics—release and dissipation—using the artificial endocrine system proposed by Neal and Timmis (2003)—equations 4 and 5. The first element is a gland (g) that releases the hormone as a function of the intensity of the external stimulus predator ($s_{predator}$) and a constant releasing rate β_g:

$$r_g = \beta_g \cdot s_{predator} \quad (4)$$

The concentration of hormone suffers two opposite forces over time: it increases with the release of hormone by the gland, and dissipates or decays over time at a constant rate γ_g:

$$c(t+1)_g = max[(c(t)_g \cdot \gamma_g) + r_g, 100] \quad (5)$$

Note that we constrain the hormonal concentration to a maximum of $c_g = 100$. In this first implementation, we have decided to do so in order to keep more

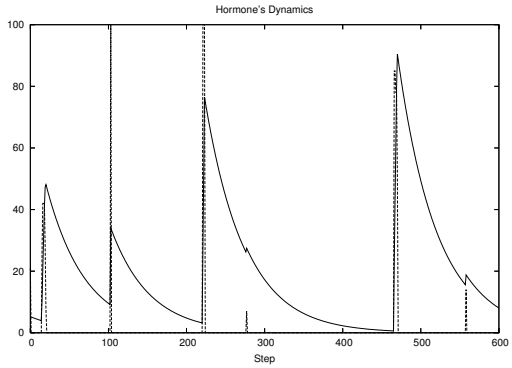


Figure 5: Hormonal level during a total of 600 steps. The dotted line represents the stimulus predator $s_{predator}$ (scaled between 0 and 100 to be shown in comparison with the hormonal level). When the predator is detected ($s_{predator} > 0$) there is hormone release proportional to the intensity of the stimulus. After any release the hormone concentration decays with time.

control on the hormone's dynamics and thus facilitate the analysis of results. Figure 5 shows the hormone's dynamics given a release rate of $\beta_g = 0.25$, and a decay rate of $\gamma_g = 0.98$. Those values were set by trial and error prior the experiments. We can observe how the hormone is released when the predator is detected ($s_{predator} > 0$) and how it decays with time.

In this implementation, the hormone increases the perception of the *integrity deficit* ($d_{integrity}$): the higher the hormone concentration the higher the reading of the $d_{integrity}$ interoceptor:

$$d^{new}_{integrity} = max(d_{integrity} + \delta_g \cdot c_g, 1) \quad (6)$$

Factor δ_g determines how susceptible to hormonal modulation the $d_{integrity}$ interoceptor is. We use $\delta_g = 0.005$, which implies that the level of perceived $d_{integrity}$ is increased by 0.5 when the hormonal concentration is maximum ($c_g = 100$). In other words, although the level of *integrity* is at its ideal value ($d_{integrity} = 0$), the interoceptor will perceive a level of 0.5 if the hormone concentration is maximum. Note that there is a constraint to avoid the level of *integrity* deficit to be perceived beyond the maximum possible value ($d_{integrity} = 1$).

Figure 6 shows the architecture with the new hormonal mechanism, that increases the perceived *integrity* deficit when the *predator* is nearby.

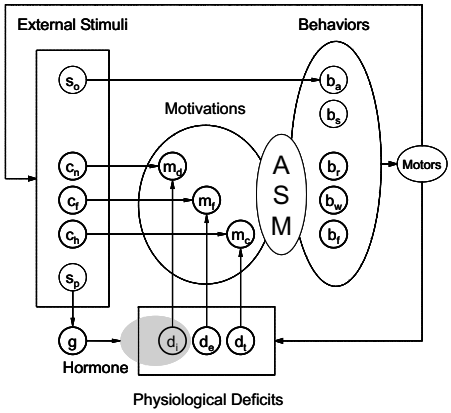


Figure 6: Proposed hormone-like modulatory mechanism for the action selection architecture. The hormone is secrete as a function of the stimulus predator ($s_{predator}$); its concentration decays over time, and modifies the perception of *integrity* deficit by the action selection architecture.

4.2 Experiments

We tested the robot during 16 runs of 1600 steps each. This means that each architecture was tested for almost two hours in our H3RP.

The prey robot presented higher stability and viability levels when equipped with the hormone-like mechanism. Please refer to (Avila-García and Cañamero, 2004a) for a more detailed quantitative analysis in terms of viability indicators and activity cycles. In this paper we focus on the difference in outer behavior between the prey with and without hormone. With hormonal mechanism, the increment in the execution time of *recover*-related (appetitive and consummatory) activities is statistically highly significant, i.e., the robot spends more time looking for the nest and recovering *integrity* in it. Other important phenomenon is the interruption of ongoing consummatory *feeding* or *warming-up* activities. When the robot is under the effect of the hormone it will abandon the resource and come back to the nest before the motivation has been satiated. The prey robot, when equipped with hormonal mechanism, presents statistically higher levels of interruption of ongoing *feeding* or *warming-up* activities.

Prey animals use unconditioned and conditioned predator cues to assess risk of predation. Curio (1993) suggests that a less studied but equally important feature is their ability to perceive risks in the absence of such cues. One example is the *risk of permanence*, or maintained levels of vigilance after predator's disappearance. Our long-term hor-

monal modulation of action selection may be seen as predation risk assessment in the absence of predator cues. Risk of predation strongly influences animal decision-making, for example, when to feed, where to feed, vigilance, or the use of nest (Lima and Dill, 1990). Risk of predation has been proposed to increase the animal's level of "apprehension," that is the reduction in attention to other activities—e.g. foraging—as a result of increasing the time executing defense-related activities—e.g. vigilance or refuge use— (Kavaliers and Choleris, 2001). Our hormone-like mechanism may be seen as increasing the level of apprehension of the prey robot after short-term predator exposure. This is reflected in an increment of the *recover* execution time at the cost of the other two activities—*feed* and *warmup*.

5 Conclusion

It is widely accepted that exposure to certain stimuli results in secretion of neurohormones–hormones and neurotransmitters—that have a long-term effect on antipredator and conspecific defensive/aggressive behavior in animals (Blanchard et al., 1998, 2001).

In this paper we first show how hormonal modulation of the perception of external stimuli (exteroceptors) can adapt the same architecture to new environmental circumstances, where the robot instead of being alone in an environment must compete with another "conspecific" for the same resources. The robot with hormonal modulation performs better than the one without it; moreover, it shows some emergent behavioral phenomena that could be interpreted by an external observer as agressive/defensive behavior.

In the second part of the paper we show how the temporal dynamics of hormones helps the same action selection architecture to adapt to a new dynamic environment. The hormonal modulation acts this time on the interoceptor of one of the physiological variables, making the action selection process more sensitive to its level. With the hormonal modulator the robot performs better, and some new behavioral phenomena emerge that could be interrelated as "fleeing" behavior and "apprehension" in a standard prey-predator scenario.

We suggest that such a modulatory hormonal systems improves the *structural coupling* between an animat and its environment in a way that is different from other mechanism such as learning or evolution, for which "past solutions" are "overwritten" by new ones. Hormonal modulation can change the functioning of the same action selection architectures, by simply acting at its sensory input, to produce adaptive behavior to any of the environmental conditions of the animat's niche.

Neuromodulation has already been related to emotions. (Fellous, 2004) proposes emotions as neuromodulatory patterns of a neural substrate, and *context dependent computation* as one of their possible functions in robotics. It is interesting to note that our two instances of hormonal modulation effectively produces a context dependent action selection, where the context is the risk of death due to, in the first place, the presence of a competitor and, in the second case, the presence of a predator.

Acknowledgments

Orlando Avila-García is funded by a research scholarship of the University of Hertfordshire. This research is partly supported by the Network of Excelence HUMAINE, EU-FP6-IST contract 507422.

References

R.A. Arkin. Homeostatic control for a mobile robot: Dynamic replanning in hazardous environments. *Journal of Robotic Systems*, 9(2):197–214, 1992.

W.R. Ashby. *Design for a Brain*. London: Chapman & Hall, 1952.

J-P. Aubin. Elements of viability theory for animat design. In *From Animals to Animats 6: Proceedings of the Sixth Intl. Conf. on Simulation of Adaptive Behavior (SAB00)*, pages 13–22. Cambridge, MA: The MIT Press, 2000.

O. Avila-García and L. Cañamero. A comparison of behavior selection architectures using viability indicators. In *International Workshop on Biollogically-Inspired Robotics: The Legacy of W. Grey Walter*, pages 86–93, Bristol HP Labs, UK, August 2002.

O. Avila-García and L. Cañamero. Long-term hormonal modulation of perception in action selection within a prey-predator robotic scenario. *Submitted to Adaptive Behaviour*, December 2004, December 2004a.

O. Avila-García and L. Cañamero. Using hormonal feedback to modulate action selection in a competitive scenario. In *From Animals to Animats 8: Proceedings of the Eight Intl. Conf. on Simulation of Adaptive Behavior (SAB04)*, pages 243–252. Cambridge, MA: The MIT Press, 2004b.

O. Avila-García, L. Cañamero, and R. Boekhorst. Analyzing the performance of "winner-take-all" and "voting-base" action selection policies within the two-resource-problem. In *Proceeding of the Seventh European Conference in Artificial Life (ECAL03)*, pages 733–742. Springer-Verlag, August 2003.

D.C. Blanchard, G. Griebel, R.J. Rodgers, and R.J. Blanchard. Benzodiazepine and serotonergic modulation of antipredator and conspecific defense. *Neuroscience and Biobehavioral Reviews*, 22(5):597–612, 1998.

R.J. Blanchard, C.R. McKittrick, and D.C. Blanchard. Animal models of social stress: Effects on behavior and brain neurochemical systems. *Physiology & Behavior*, 73:261–271, 2001.

B. Blumberg. Action selection in hamsterdam: Lessons from ethology. In *From Animals to Animats 3: Proceedings of the Third Intl. Conf. on Simulation of Adaptive Behavior (SAB94)*, pages 108–117. Cambridge, MA: The MIT Press, 1994.

R.A. Brooks. A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, RA-2: 14–23, April 1986.

L. Cañamero. Modeling motivations and emotions as a basis for intelligent behavior. In W.L. Johnson, editor, *Proceeding of the First Intl. Conf. on Autonomous Agents (Agents97)*, pages 148–155. New York: ACM Press, 1997.

E. Curio. Proximate and developmental aspects of antipredattor behavior. *Adv. Study Behav.*, 22:135–238, 1993.

J-M. Fellous. From human emotions to robot emotions. In E. Hudlicka and L. Cañamero, editors, *Proceedings of Architectures for Modeling Emotion: Cross-Disciplinary Foundations. 2004 AAAI Symposium*, pages 37–47, Stanford, California, March 2004.

A. Guillot and J-A. Meyer. Computer simulations of adaptive behavior in animats. In N. Thalman and G. Thalman, editors, *IEEE Proc. on computer Animation '94*. IEEE Computer Society Press. Silver Spring, MD, 1994, 1994.

R.M. Harris-Warrick, F. Nagy, and M.P. Nusbaum. Neuromodulation of the stomatogastric networks by identified neurons and transmitters. In R.M. Harris-Warrick, E. Marder, A.I. Selverston, and M. Moulins, editors, *Dynamic Biological Networks*, chapter 3, pages 87–137. The MIT Press, 1992.

M. Kavaliers and E. Choleris. Antipredator responses and defensive behavior: Ecological and ethological approaches for the neurosciences. *Neuroscience and Biobehavioral Reviews*, 25:577–586, 2001.

E.A. Kravitz. Hormonal control of behavior: Amines and the biasing of behavioral output in lobsters. *Science*, 241 (4874):1175–1781, September 1988.

I.B. Levitan and L.K. Kaczmarek. *The Neuron: Cell and Molecular Biology*. Oxford University Press, 1997.

S.L. Lima and L.M. Dill. Behavioral decitions made under the risk of predation: a review and prospectus. *Can. J. Zool.*, 68:619–640, 1990.

P. Maes. A bottom-up mechanism for behavior selection in an artificial creature. In *From Animals to Animats: Proceedings of the First Intl.Conf. on Simulation of Adaptive Behavior (SAB90)*, pages 238–246. Cambridge, MA: The MIT Press, 1991.

P. Maes. Modeling adaptive autonomous agents. In C.G. Langton, editor, *Artificial Life: An Overview*, pages 135–162. Cambridge, MA: The MIT Press, 1995.

D. McFarland, editor. *Motivational Control Systems Analysis*. London: Academic Press, 1974.

D. McFarland. *Animal Behaviour*. Addison Wesley Longman Limited, 3rd edition, 1999.

J-A. Meyer. The animat approach to cognitive science. In H. L. Roitblat and J-A. Meyer, editors, *Comparative Approaches to Cognitive Science*, chapter 2, pages 27–44. The MIT Press, 1995.

M. Neal and J. Timmis. Timidity: A useful emotional mechanism for robot control? *Informatica*, 27:197–204, 2003.

T.W. Robbins and B. J. Everitt. Motivation and reward. In M.J. Zigmond, F.E. Bloom, S.C. Landis, J.L. Roberts, and L.R. Squire, editors, *Fundamental Neuroscience*, chapter 48, pages 1245–1260. Academic Press, 1999.

R.M. Sibly and D. McFarland. A state-space approach to motivation. In D. McFarland, editor, *Motivational Control Systems Analysis*, chapter 5, pages 213–250. London: Academic Press, 1974.

E. Spier and D. McFarland. Possibly optimal decision making under self-sufficiency and autonomy. *J. Theor. Biol.*, 189:317–331, 1997.

L. Steels and R. Brooks. *The Artificial Life Route to Artificial Intelligence: Building Situated Embodied Agents*. New Haven: Lawrence Erlbaum Associates, 1995.

N. Timbergen. *The Study of Instinct*. Clarendon Press, 1951.

F. Toates. *Motivational Systems*. Cambridge Univ. Press, 1986.

T. Tyrrell. *Computational Mechanism for Action Selection*. PhD thesis, Centre for Cognitive Sciences, University of Edinburgh, online at: http://www.cs.bham.ac.uk/ sra/People/Stu/Tyrrell, 1993.

S.W. Wilson. Knowledge growth in an artificial animal. In *Proceedings of the First Intl. Conf. on Genetic Algorithms and their applications*, pages 16–23. Hillsdale, NJ. Lawrence Erlbaum, 1985.

Personality and Learning in Robots. The Role of Individual Motivations/Expectations/ Emotions in Robot Adaptive Behaviours.

Barbara Caci[*]
[*]Dipartimento di Psicologia
Università di Palermo
Viale delle Scienze, Edificio 15
90128, Palermo, Italy
Phone: +39 091 7028420
FAX +39 091 7028430
bcaci@unipa.it

Maurizio Cardaci[*†]
*Dipartimento di Psicologia and
[†]CITC -Università di Palermo
Viale delle Scienze, Edificio 15
90128, Palermo, Italy
Phone: +39 091 7028415
FAX +39 091 7028430
cardaci@unipa.it

Antonio Chella[#†°]
[#]DINFO and
[†]CITC and
[°]ICAR- CNR, Palermo
Università di Palermo,
Viale delle Scienze, Edificio 6
90128, Palermo, Italy.
Phone: +39 091 6615239
FAX +39 091 488452
chella@unipa.it

Antonella D'Amico[*]
[*]Dipartimento di Psicologia and
[†]CITC -Università di Palermo
Viale delle Scienze, Edificio 15
90128, Palermo, Italy
Phone: +39 091 7028420
FAX +39 091 7028430
adamico@unipa.it

Ignazio Infantino[°]
[°]ICAR- CNR, Palermo
Viale delle Scienze, Edificio 11
90128, Palermo, Italy.
Phone: +39 091 238262
FAX +39 091 6529124
infantino@pa.icar.cnr.it

Irene Macaluso[#]
[#]DINFO, Università di Palermo,
Viale delle Scienze, Edificio 6
90128, Palermo, Italy.
Phone. +39 091 6615239
FAX +39 091 488452
macaluso@csai.unipa.it

Abstract

The present paper is aimed to study the influence of different personality factors, implemented via complex software architecture, in the exploration of environment by mobile robots like RWI-B21. We adopted the *social cognition* framework (e.g. Rotter, 1960; Rotter, Chance, & Phares, 1972; Bandura, 1977) that considers the individuals' motivated and emotionally oriented behaviours as the result of their cognitive evaluations of the environment demands and of their own capabilities to cope it. In this framework, we present two robots provided with an internal vs. external locus of control. Our robot architecture, that integrates motivations, emotions and symbolic knowledge representation by means of a rich and expressive conceptual area where affective computing takes place, is based on a constant matching between the expected results of a goal-oriented action and its real outcomes. The psychological evaluation of the success/failure of the goal-oriented behaviour is modulated by the robot locus of control that regulates the expectancy updating values and the mood state used by robots in exploring the environment. We present some experiments of the proposed architecture, performed on RWI-B21 robot assigned with a navigation task.

1 Introduction

The present paper is aimed to study the influence of different personality factors, implemented via complex software architecture, in environmental tasks executing by mobile robots like RWI-B21.

To this aim, we adopted the *social cognition* framework that considers the individuals' motivated and emotionally oriented behaviours as the result of their cognitive evaluations of the environment demands and of their own capabilities to cope it.

According to the *social learning theory* (Rotter, 1960; Rotter, Chance, & Phares, 1972; Bandura, 1977) the individuals' motivated behaviours arise

from the relatively consistent and characteristic ways in which a person represents, decides and *expects* to reach his/her goals evaluating the events according his/her success/failure perceptions. Therefore, the human behaviour is goal-oriented and the chance that a given behaviour occurs is a function of two combined factors: the first is the expectation that a particular behaviour will obtain a reward; the second is the perceived value of this reward by the individual. The perception of reward is modulated by the individuals' *locus of control* (e.g. Rotter, 1966; 1975; Strickland, 1978; Cardaci, 1988; Marshall, Collins, & Crooks, 1990; Lefcourt, 1991; McLaughlin, & Saccuzzo, 1997). People with an *internal locus of control* expect to be personally responsible of their outcomes (in terms of success/failure); internality is associated both with high motivational levels and high expectations of obtaining the reward. People with an *external locus of control* attribute their outcomes to a variety of external causes as luck, fate, others, and so on; externality is associated both with low motivational levels and low expectations of obtaining the reward. In this perspective, when a "junctures of plans" occurs (Oatley & Johnson-Laird, 1987), or in other words, when people feel a smoothing or strong mismatching between the expected result of a planned behaviour and the real outcome, their locus of control modulates the emotive appraisals of the event and the following generation of coping strategies (Lazarus, 1966; Frijda, 1986). Rotter (1966; 1975) found strong support for the hypothesis that the locus of control influences the emotionally oriented behaviours, inducing internal people to believe that they can control their own destinies by personal efforts and resources (e.g. Rotter, 1966; McLaughlin, & Saccuzzo, 1997) and, conversely, external people to believe that results are not under their control.

2 The Architecture

In this framework, we developed a robotic architecture that integrates motivations, emotions and symbolic knowledge representation in performing different environmental tasks (see Figure 1). It allows two robots, which differ each other about their "personality" and in particular in their locus of control, to represent and to explore the environment by means of a rich and expressive conceptual area, based on the theory of conceptual spaces (Gärdenfors, 2000). Conceptual spaces provide a principled way for relating high level, linguistic formalisms, with low level, unstructured representation of data. A conceptual space (CS) is a metric space whose dimensions are related to the quantities processed by the robot sensors. Some dimensions are related to object's shape, while other dimensions are related to the displacement in space of the moving 3D shape. We adopt the term *knoxel* to denote a point in the conceptual space. In order to account for the perception of dynamic scenes, we choose to adopt an intrinsically dynamic conceptual space (Chella, Frixione & Gaglio, 2000). The robot CS is generalized to represent moving and interacting entities: the generic motion of an object is represented in its wholeness, rather than as a sequence of static frames (Marr &Vaina, 1982).
The dynamic conceptual space lets the agent to imagine possible future interactions with its environment. The evolution of the scene is represented as a sequence of sets of *knoxel* that is imagined and simulated in the conceptual space before the interaction really happens in the real world expectations. The link between the conceptual space representation of the robot and the robot behaviour system is described in details in (Chella, Gaglio & Pirrone, 2001).

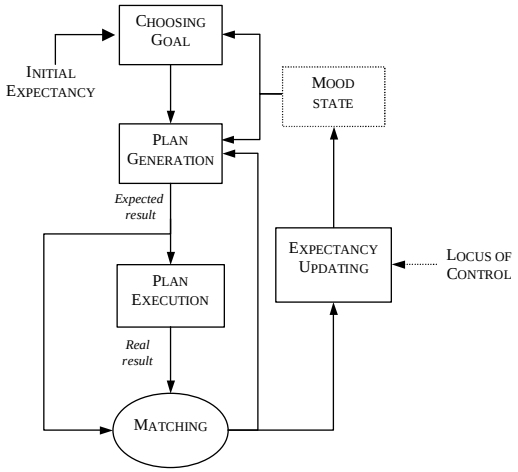


Figure 1: The robot architecture

Both robots, indeed, are provided with a similar *initial expectancy* used for choosing a goal; in the following *plan generation* phase, they generate a plan of actions in order to produce the expected situations. This level of representation can be considered as the robot mental imagery and, as previously stated, is represented in the conceptual space as a sequence of sets of *knoxel*. During *plan execution*, the robot perceives the environment and it is able to compare the *expected results* of its actions, as they were simulated during plan generation, and the *real results* according to its current perceptions.

The differences between the two robots' personalities arise in the *expectancy updating* phase that is modulated by the robots' *Locus of Controls*. The updating of the expectancy value, that depends on the result of the *matching*, simulates the internality of the robot; the updating of the expectancy value in a random way simulates the externality of the robot. The expectancy updated value is used by both robots to increment/decrement their *mood states*: the internal robot experiments a range of mood states from a positive one in the case of total matching to a particularly "depressed" one in the case of mismatching. In the first case, the robot will behave more persistently in the achievement of goals of increased difficulty level; in the second case, the robot experiences a sort of "learned helplessness" (Seligman, 1972) that bring it to choose goals of decreased difficulty level. On the contrary, the *external* robot experiments its mood states in a random way both in the case of matching and mismatching, independently of the behavioural outcomes. The mood states affects both the execution speed of the following behaviours and, as shows in Figure 2, the region surrounding the set of knoxels representing the perceived and expected situations. The changes in the conceptual space influences the new plan generation and/or the choosing of new goals.

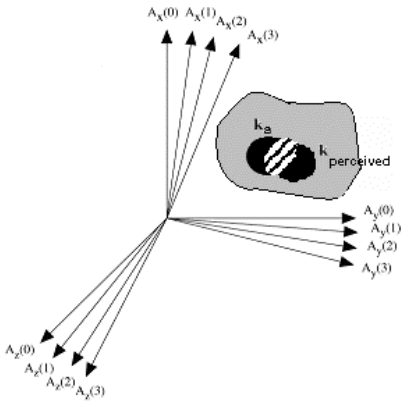


Figure 2: A pictorial representation of the conceptual space of an internal robot: in case of partial matching the robot experiments a "grey" mood state.

3 Experimental Results

The proposed architecture was tested on an RWI-B21 robot assigned with a navigation task in a known environment. In order to measure the expectancy updating value, we simulated various success/failure situations experienced by the internal and external robots living in the same environment (we fixed to 0.5 the initial expectancy value for both robots). In Table 1, we reported the expectancy updating values associated with various matching sequences.

Table 1: Expectancy updating values corresponding to various matching sequences.

Matching Sequences	**Expectancy updating value**		
	Internal Robot	External Robot	Difference I-E
1; 1	12.5	4.0	8.5
0.8; 1	11.3	1.7	9.6
0.5; 1	9.5	-0.3	9.8
0.5; -1	-2.5	-3.7	1.2
1; -1	0.5	2.6	-2.1
1; -0.1	6.5	-0.9	7.4
-1; 1;	0.5	1.2	-0.7
-1; -1	-11.5	-2.1	-9.4

[-1; 1] ≡[Total Mismatching ; Total Matching]

As hypothesized, the expectancy updating value is quite different for the internal/external robots. For example, in Table 1, the first row represents the result of success/success situations (six iteration): the *internal* robot reaches an expectancy updating value of 12.5, while, the *external* robot reaches an expectancy updating value of 4. Such values affect the robots' mood in a sensible different way. In the third column is possible to observe that, depending on each success/failure sequence (e.g. 1,1; 0.5, 1; -1, -1), the difference between the I-E robots is more or less amplified.

4 Conclusion

These results indicate that, as observed in human beings (e.g. Rotter, 1966; 1975), the internal robot is more persistent and motivated in reaching its goals in highly predictable environments, but his internality becomes a "depressing" factor in more dynamic and unpredictable ones. On the contrary, the external robot doesn't vary its expectancy value depending on the environmental context; this feature allows him to be less emotionally vulnerable than the internal robot in dynamic and unpredictable environments and less persistent and motivated in predictable ones.

The obtained results demonstrate the utility of multidisciplinary research in the field of Affective Computing. In particular, the implementation of psychological theories in robotics agents may both validate the specific models of personality used and

provide human-like personalities to traditionally rational artificial agents.

References

A. Bandura, *Social Learning Theory*. General Learning Press, New York, 1977.

M. Cardaci, Studi sul Locus of Control, *Contributi del Dipartimento di Psicologia*, Università degli Studi di Palermo, 1988.

A. Chella, M. Frixione, S. Gaglio, Understanding dynamic scenes, *Artificial Intelligence*, Vol.123: pp. 89-132, 2000.

A. Chella, S. Gaglio, R. Pirrone, Conceptual Representations of Actions for Autonomous Robots, *Robotics and Autonomous Systems*, Vol. 34: pp. 251–263, 2001.

N. H. Frijda, *The Emotions*, Cambridge University Press, New York, 1986.

P. Gärdenfors, *Conceptual Spaces*. MIT Press, Bradford Books, Cambridge, MA., 2000.

R. S. Lazarus, *Psychological Stress and the Coping Process*, McGraw-Hill, New York, 1966.

H. M. Lefcourt, Locus of Control, in Robinson J.P., Shaver P.R., Wrightsman L.S, (Eds.). *Measures of Personality and Social Psychological Attitudes*, Vol. 1**:** pp. 413-499. Academic Press, Inc., San Diego, CA., 1991.

D. Marr, L. Vaina, Representation and recognition of the movements of shapes, *Proceeding of Royal Society London*, Vol. B, No 214: pp.501-524, 1982.

G. N. Marshall, B. E. Collins, & V. C. Crooks, A comparison of two multidimensional health locus of control instruments. *Journal of Personality Assessment*, Vol. 54: pp. 181-90, 1990.

S. C. McLaughlin, & D. P. Saccuzzo, Ethnic and gender differences in locus of control in children referred for gifted programs: The effects of vulnerability factors. *Journal for the Education of the Gifted*, Vol. 20: pp. 268-83, 1997.

K. Oatley, & P. N. Johnson-Laird, Towards a cognitive theory of emotions. *Cognition and Emotion*, Vol. 1: pp. 29-50, 1987.

J. B. Rotter, Some implications of a social learning theory for the prediction of goal directed behavior from testing procedures. *Psychological Review*, Vol. 67: pp. 301-316, 1960.

J. B. Rotter, Generalized expectancies for internal versus external control of reinforcement. *Psychological Monographs*, 80, 1966.

J. B Rotter, Some problems and misconceptions related to the construct of internal versus external control of reinforcement. *Journal of Consulting and Clinical Psychology*, Vol. 43: pp. 56-67, 1975.

J. B. Rotter, J. E. Chance, & E. J. Phares, *Applications of a social learning theory of personality*, Holt, Rinehart & Winston, New York, 1972.

M.E.P. Seligman, Learned helplessness, *Annual Review of Medicine*, Vol. 23: pp. 407-412, 1972.

B. R. Strickland, Internal-external expectancies and health-related behaviors. *Journal of Consulting and Clinical Psychology*, Vol. 46 : pp. 1192-1211, 1978.

A model of emotional influence on memory processing

Philippe Chassy[*]
[*]Centre for Cognition and NeuroImaging
Brunel University, Uxbridge Middlesex, UB8 3PH
Philippe.Chassy@Brunel.ac.uk

Fernand Gobet[†]
[†]Centre for Cognition and NeuroImaging
Brunel University, Uxbridge Middlesex, UB8 3PH
Fernand.Gobet@Brunel.ac.uk

Abstract

To survive in a complex environment, agents must be able to encode information about the utility value of the objects they meet. We propose a neuroscience-based model aiming to explain how a new memory is associated to an emotional response. The same theoretical framework also explains the effects of emotion on memory recall. The originality of our approach is to postulate the presence of two central processing units (CPUs): one computing only emotional information, and the other mainly concerned with cognitive processing. The emotional CPU, which is phylogenetically older, is assumed to modulate the cognitive CPU, which is more recent. The article first deals with the cognitive part of the model by highlighting the set of processes underlying memory recognition and storage. Then, building on this theoretical background, the emotional part highlights how the emotional response is computed and stored. The last section describes the interplay between the cognitive and emotional systems.

1 Introduction

Intensive research in neuroscience has established a close link between emotion, cognition, and action. In recent years, researchers in artificial intelligence and robotics have attempted to build artificial systems where emotional and motivational mechanisms modulate cognitive mechanisms (e.g., Pfeifer & Scheier, 1999). However, these attempts have so far been directed at modelling animal behaviour and/or relatively simple tasks, and did not use current knowledge of the human brain. We suggest that, in order to understand how emotion, motivation, cognition and action interact in complex systems it is necessary to use the information provided by current research in neuroscience.

The aim of this paper is to propose a model linking emotion and cognition that is based on recent developments in neuroscience. In particular, we want to identify the learning and memory mechanisms that enable memories to be influenced by emotions. We also discuss how the proposed mechanisms relate to chunking, a mechanism that has been shown to be central to cognition from simple animals to human experts. While we do not present a computer or robotic implementation of these mechanisms, we believe that they will be of interest to researchers building such implementations and others attending this symposium. In particular, we hope that the ideas presented in this theoretical paper will entice colleagues to build computer or robotic systems that will allow testing them empirically.

The first section presents a model of memory processing. The model describes the mechanisms behind object recognition and storage, and spells out the role of the cognitive central processing unit (CPU). In the second section, we present a model of emotional processing. The objective is to show how an affect is generated from recognised objects and to make clear the relation between the retrieval of stored information and that of emotional response. In this section, we also introduce the concept of an emotional CPU, which computes the emotional response. Finally, we present an integrative model that consists of the two previous components. Thus, a direct link is made between memory storage and emotional processing. The interaction between the two CPUs is crucial, as it is the means by which the emotional responses are actually influencing cognitive processing.

2 The cognitive system

2.1 Bottom-up processing

The recognition of an object is done when a particular configuration of neurons coding for the object and its properties is activated. Such distributed neural networks, which activate a small number of modules concerned with different types of informa-

tion (spatial, auditory, etc…), are known as *cell assemblies* (Sakurai, 1999).

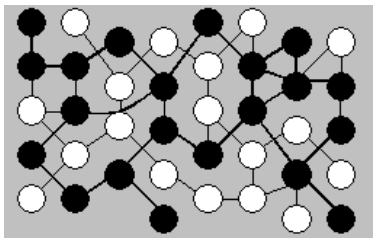


Figure 1. A cell assembly (CA). Cells possibly located in distinct parts of the brain are activated jointly.

A cell assembly is distributed in distinct brain areas. The cells forming a cell assembly are closely interconnected (bold links in the Figure 1): the activation of a neuron belonging to the assembly is likely to activate the other neurons of the assembly. This propagation law makes the CA act as a unit: when a subpart of the CA is activated (e.g., the property of an object is recognised), the spreading activation is very likely to activate all the CA, so that the object is recognised even if some of the information is missing (e.g., when an object is partially hidden).

As soon as an object is recognised, the neural coordinates serve as a pointer for *short-term memory (STM*) storage. It is likely that top-down inhibitory control processes are carried out at this stage, specifying whether or not the recognised object is of interest and should accordingly be stored in STM.

2.2 The functions of the cognitive CPU

When a CA is activated (i.e., an object is recognised), a pointer that codes for this CA is stored in STM. As each recognised object activates a CA, the role of the cognitive CPU is to compute an overall representation of the external milieu by connecting the active CAs. This computation is likely to have two main consequences:

(i) To activate new pathways *between* the cell assemblies.
(ii) To modify the dynamics *within* each cell assembly. The spreading activation of any cell assembly is now an input for all others.

The result of this computation is the emergence of a new *dynamic neural network* (DNN) that represents the external milieu. The DNN is maintained active by CPU processing. Unlike CAs, the DNN needs an active control to remain active.

2.3 Consolidation: Long-term memory

The question now arises as to how the DNN may be encoded more permanently. In line with this theoretical framework, *long-term memory (LTM*) storage refers to the processes that change the neural network activity computed by the cognitive CPU into one consolidated memory trace. Neuroscience is bringing converging evidence supporting the view that the medial temporal lobe mediates information storage in the sensory cortices by generating structural changes via mechanisms like long-term potentiation (Kandel, Schwartz, & Jessell, 2000). We suggest that, following the computation of a new DNN (see previous section), the CPU supervises its consolidation into durable memory traces. Such consolidation is done through the synthesis of new synapses consolidating the new connections (Kandel et al., 2000). As a result of the consolidation of memory traces, a new durable CA emerges, and thus a new object can be recognised. We note that the new CA encompasses the previous CAs coding for the individual objects. This process of building new patterns by accretion of previous ones is known as chunking (Gobet et al., 2001).

Chronologically, the chain of reactions leading to the emergence of a new CA is initiated by bottom-up processing which generates the recognition of objects. As soon as recognition is done, the CPU computes a dynamical network that represents the situation in STM. Thanks to the consolidation process, the dynamical network evolves into a structural network. This leads to the emergence of a new cell assembly that could be activated by discrimination of entrant stimuli.

So far, we have presented a theoretical framework for memory recognition and storage. The key component is the cognitive CPU, the function of which is twofold: first, to compute the representation of the external milieu based on the recognised objects and, second, to consolidate the new computed representation in memory. In the next section, we turn our attention to emotional processing.

3 The emotional system

3.1 The functions of emotions

The information processed by the cognitive system, although useful for survival, lacks any utility value. However, a living system needs to know the utility of the objects in the environment in relation to a given task. Emotions, which play this role, have been defined by Rolls (2003) as "*states elicited by rewards and punishers*" (p. 552). A reward is what a living system is ready to work for, and a punisher is

what it tries to avoid. Emotions are thus goal-directed, and values depend on the goal (the motivator). For example, when one is cooking, the goal is to prepare good food (motivator). Each element taking part in the action of cooking has its own utility (emotional value: reward or punisher). In general, emotions structure the environment by tagging objects with an emotional value relative to a goal (the motivator). The emotional physiological response consists in preparing the body for action (Frijda, 1993). What remains to be spelled out is how the emotional system tags information from the environment, and how this influences high-level cognitive processes.

3.2 Emotional processing

A key issue in emotion research is how a neutral stimulus comes to generate emotional responses with experience. A good example from this field of research is LeDoux's (2002) model of fear conditioning, which links emotion and cognition in the simplest way: a stimulus, previously neutral to the animal, is paired to an emotional response. This model of fear learning illustrates that emotional responses are stored, and that they can be linked to stored objects. As soon as the object is recognised, the emotional response is retrieved. Fear is an unlearned punisher, thus likely to be genetically coded (Rolls, 2003). But some emotions are experience-dependent and thus ontologically built. Therefore, we need to provide an explanation of how new emotions are computed from previous emotional responses. This necessary flexibility suggests the existence of a system regulating emotional responses.

A living system has thus to be able to process emotions that do not induce automatic responses. In addition, different objects are likely to induce different emotions, giving rise to internal conflicts. To deal with these issues, we propose a model where three steps of information processing are postulated. In the first step, any recognised object in a scene activates its associated emotional response. In the second step, the emotional CPU computes an overall emotional response in the same way as the cognitive CPU computes an overall representation. That is, the emotional responses are physiological levels that encode objects utility values; in order to integrate the emotional responses, the CPU sums the utilities of all objects. In the third and final step, the computed emotion is felt by the individual as related to the representation of the external milieu.

This model, which spells out how recognised objects can retrieve different emotions, explains how the living system generates new emotions based on the emotional responses of the recognised objects. But the living system also needs a way to combine the overall emotional response with the representation computed by the cognitive CPU. This issue is addressed in the next section.

4 Cognition and emotion

4.1 Empirical evidence on the influence of emotion

There is substantial empirical evidence supporting a close relationship between the cognitive and emotional systems, and two examples will suffice here. Erk et al. (2003) demonstrated that non-emotional objects are better encoded in memory when they have been associated to a strong enough emotional clue. That is, the emotional context modulates the encoding of memories with no prior emotional value. Kilpatrick and Cahill (2003) showed emotionally loaded film clips to their participants, and, using neuroimaging techniques, showed that the amygdala, a brain structure related to emotional processing, influences the hippocampus, a brain structure related to memory consolidation.

4.2 Emotion and memory storage and retrieval

The emotional and cognitive processes are both parallel and serial (see Figure 2).

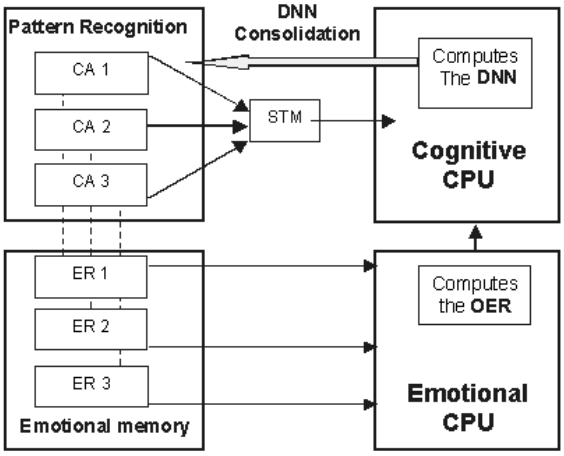


Figure 2. The information flow: a summary

In order to clarify them we divide the chain of actions in two steps. The first step is the computation of both the representation of the external milieu by the cognitive CPU, and the computation of the emotional response by the emotional CPU. The second step, explains how the two CPUs cooperate in order to link and consolidate the dynamic neural network

representing the external milieu to the computed emotional response.
Figure 2 also describes the two processes that are running in parallel: one cognitive, the other emotional. The cognitive process was explained in the first section of the paper. The emotional process is divided in two steps. Firstly, each active CA induces the retrieval of its associated *emotional response (ER1, ER2, ER3*). Secondly, the emotional CPU computes an *overall emotional response (OER)* by combining the retrieved emotional responses. The OER is the emotional utility value of the representation coded by the DNN. The OER is at the origin of the emotional influence of memory as it modulates the speed of the DNN consolidation: The more intense the OER, the more facilitated the consolidation of the DNN will be. In addition to this, the cognitive CPU consolidates the link between the DNN and the EOR. As a result of this process, the DNN is associated to the computed emotional response.

In summary, by means of the consolidation process, the cognitive CPU has created a new representation and has linked it to a new emotional response. When the representation is activated again in the future, the corresponding emotional response is retrieved.

5 Conclusions

The model provides an explanation of how recognised objects generate a representation of the external milieu and how this representation is turned into a structural feature of the system. In doing so, we put forward a biologically valid explanation for the influential concept of chunking (Gobet et al., 2001). The model also explains how emotional information is linked to objects stored in memory. Tagging allows the system to retrieve the value of an object upon its recognition. The computations that follow ensure the necessary flexibility of high-level cognition. Both for memory and cognition, we have postulated that a CPU played a key role in controlling how cell assemblies are combined together so that they can be recognised as a unit in the future.

Testing this theory will require a combination of empirical and theoretical research. That is, it will be necessary to develop computer programs or autonomous robots whose parameters will be set by both biological and psychological data. Then, experiments should be run to test whether agents would perform better in complex and dynamic environments with the presence of the emotional mechanisms described in this paper. If supported empirically, our theory would provide a powerful conceptual framework for computer science and robotics, as it would offer an explanation of how emotions help humans to structure their perceptual space.

There is no doubt that understanding—and, in a computational model or robot, controlling—the dynamics of cell assembly raises serious theoretical and practical problems. But then, the implicit message of this paper is perhaps that complex processing is necessary when emotions enter the scene and, in corollary, that there are limits in what simple autonomous agents can do without emotions modulating the way they perceive and process information.

References

Erk, S., Kiefer, M., Grothe, J., Wunderlich, A.P., Spitzer, M., & Walter, H. (2003). Emotional context modulates subsequent memory effect. *NeuroImage, 18*, 439-447.

Frijda, N. (1993). Moods, Emotions Episodes, and Emotions. In H. Lewis. & J.H. Hairland (Eds). *Handbook of Emotions.* New York: Guilford press.

Gobet, F., Lane, P.C.R., Croker, S., Cheng, P.C.H., Jones, G., Oliver, I., & Pine, J.M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences, 5*, 236-243.

Kandel, E.R., Schwartz, J.H., & Jessell, T.M. (2000). *Principles of neural science*. New-York: McGraw Hill.

Kilpatrick, L., & Cahill, L. (2003). Amygdala modulation of parahippocampal and frontal regions during emotionally influenced memory storage. *NeuroImage, 20*, 2091-2099.

LeDoux, J. (2002). *Neurobiologie de la personnalité.* Paris: Odile Jacob.

Pfeifer, R., & Scheier, C. (1999). *Understanding intelligence*. Cambridge: MIT Press.

Rolls, E.T. (2003). Vision, emotion and memory: From neurophysiology to computation. *International Congress Series, 1250*, 547-573.

Sakurai, Y. (1999). How do cell assemblies encode information in the brain? *Neuroscience and Biobehavioral Reviews, 23*, 785-796.

Analysis of the human physiological responses and multimodal emotional signals to an interactive computer

M.R.Ciceri
Lab of Communication Psychology
Catholic University of Sacred Heart
Largo Gemelli, 1 20143 Mi- Italy
{rita.ciceri}@unicatt.it

S.Balzarotti
Lab of Communication Psychology
Catholic University of Sacred Heart
Largo Gemelli, 1 20143 Mi- Italy
{stefania.balzarotti}@unicatt.it

P.Colombo
Lab of Communication Psychology
Catholic University of Sacred Heart
Largo Gemelli, 1 20143 Mi- Italy
{paola.colombo}@unicatt.it

Abstract

New theories on emotion emphasize the tight boundaries among emotion, motivation and cognition: according to them people are motivated to respond to events differently depending on how they have been appraised. The present study tries to examine what kind of emotional responses the subject is motivated to express while interacting with an artificial agent if he/she believes that this agent is able to understand his/her emotional states. For this purpose different kind of computer games were projected to elicit specific emotional appraisals and two different conditions were used: in one condition the avatar provided a simulated intelligent feedback, while in the second only guided the subject across the different tasks. Multimodal synchronized data were captured. All video tapes were codified frame by frame using The Observer 5.0 and THEME Software. This is an in-progress study and data are still under elaboration. This paper includes initial results of the analysis of non-verbal communicative signals.

Introduction

New theories of mind try to gather the continuity between cognition and action and their tight boundaries with motivation and emotion: their goal is the elaboration of unified models where these functions are merged to explain human behaviour. In this perspective, those theories suggest that mind is «embodied» and situated as it originates in the interaction with the environment (Barsalou, 1999; Lakoff & Johnson, 1999). Humans are then considered as biological agents able of acting on the environment to change it and of functionally adapting in relation to their own needs. Procedural actions always involve emotion and motivation: agents organize action sequences to achieve goals relying on the emotional and cognitive evaluation of events. Motivation, cognition and emotion are no longer considered as separated functions and all support executing actions effectively.

Theoretical knowledge suggests that for human beings emotions reflect an adaptive system and represent a crucial functional aspect to maintain a satisfactory relationship with the environment (Frijda, 1986; Scherer, 1993). As a matter of fact, emotions correspond to different forms of action tendency (e.g. avoid, approach, interrupt, change strategy, attempt, reject, etc.) and involve states of action readiness (Frijda, 1986) elicited by circumstances appraised as relevant to the subject's goals. In this perspective, they can constitute very powerful motivational factors: people are goal-driven, task-solving agents motivated to action by a complex system of emotions.

These issues open interesting questions when applied to the context of the interaction between biological and artificial agents. The increasing use of computers and machines that support the human user in any kind of task has nowadays transformed them into important and constitutive members of the physical and social environment with which people interact. For this reason, researchers from several disciplines have turned their attention to emotion and its place in the interaction between human and artificial agents, trying to understand and re-create emotion in the user interface (Picard, 1997; Picard, Vyzas, Healey, 2001; Lisetti, 2002; Lisetti, et al., 2003; Norman, 2003).

Emotional models have been proposed as a critical component of more effective human computer interaction: explicit attention to the emotional aspects aims at increasing the system performance in terms of usability and acceptance.

In this sense, emotions play an important role in the design of interfaces: they should not be considered a simple optional providing pleasantness but they represent crucial cues as they are involved in the selection, regulation and motivation to action.

1.1. Emotion and cognitive appraisal

During the last twenty years, appraisal theories have suggested that emotions are elicited and differentiated on the basis of a person's subjective cognitive evaluation of the significance of a situation (Weiner, 1986; Lazarus, 1991; Roseman, 1991; Scherer; 1993). In particular, they postulate a fixed sequence of stimulus evaluation checks: novelty and expectancy, valence, relevance and goal conduciveness, agency and responsibility, perceived control (Scherer, 2000; 2001; Van Reekum, et al., 2004).

People are then motivated to respond to events differently depending on how they have been appraised (Lazarus, Smith, 1988; Wehrle, Scherer, 1995; Gratch, Marsella, 2004).These theories suggest that emotion episodes require the individual to adapt and actively respond to the situation.

For this reason, during emotion all major functioning subsystems of organism are recruited to interact: cognition, physiological regulation, motivation, motor expression and monitoring/ feeling (Scherer, 1993; Frijda, 2001; Mesquita, Frijda, Scherer, 1997; Anolli, Ciceri, 1997; Roseman, et al, 2001; Ciceri 2002). Emotion is then described as a *process*, that is, as the dynamic time course of *constantly changing affective tuning* of organisms as based on continuous evaluative monitoring of their environment.

Cognitive criteria of evaluation influence each of the response components. In particular, great interest has focused on physiological arousal and expressive behaviour. For example, through the systematic manipulation of types of events in a computer game (Kappas, Pecchinenda, 1999; 2000; Van Reekum, et al., 2004) it has been possible to prove the influence of appraisal dimensions on physiological reactions and vocal behaviours. Appraisal criteria influenced behaviour in different ways. For example, the manipulation of intrinsic pleasantness had little impact on physiological responses while goal conduciveness was associated with relevant autonomic effects. Results show the utility of trying to manipulate emotion appraisals to measure emotional reactions and suggest that computer games offer a useful area of research to control other appraisal checks such as coping.

1.2. Emotional agents: capturing signals or communicative agreement?

In «The Media Equation» Reeves and Nass (1996) argue that human-machine interaction is inherently natural and social, so that the rules of human-human interaction apply to human-machine interaction: in many ways people seem to respond psychologically to interactive computers as they were human actors and not tools (Picard, Klein, 2002; Gratch, Marsella, 2004). Agents that show to understand emotion and behave like humans in the environment where they interact with users (such as computer games or tutoring environments) are more enjoyable and engaging.

Different computational models of the user's emotion have been developed to support the complexity of human emotion and the multi-modal richness of face-to-face communication (Bianchi-Berthouse and Lisetti, 2002; Kort, Reilly, Picard, 2001; Cowie et al., 2001). On one side theoretical knowledge on emotion has been applied to implement artificial emotional agents and design human-like autonomous agents that interact face-to-face with the user and simulate human emotional expressions (Gratch, Marsella, 2001; Braezel, 2002; Bartneck, 2001; Lisetti et al., 2004).

On the other side, a particular interest has concerned the implementing and design of interfaces able to recognize the user's emotional state from real-time capturing and processing of sensory modalities input via various media: physiological, facial and vocal signals (Lisetti, et al., 2003). This may allow the machine to adapt intelligently the running task to the recognized emotional state of the user in order to support his actions and motivation to the task itself.

The idea of a computer «sensing» the user's autonomic nervous system (ANS) activity is becoming increasingly popular, due to the recent progress in analyzing user's physiological states (Prendiger, Mayer, Mori, Ishizuka, 2003). Our work attempts to move beyond this idea of an emotional interface that attempts to interact with the user capturing and processing involuntary signals.

On one hand, the previously presented theories tell us that humans are active agents and emotion plays an important role of in the evaluation and motivation to act. In this perspective, we suggest that emotion can not be reduced to the physiological arousal: the agent relies on communicative signals that he intentionally directs to express his state and to act on the situation.

On a second hand, within the framework of the latest models that present a *two-ways* vision of communication (Bratman, 1990; O'Keefe,

Lambert, 1995; Greene, 1997; Searle, 1998; Ciceri, 2001), we assume that a crucial factor for the interaction with an emotional artificial agent and its effectiveness is the presence of a «communicative agreement». If the user is aware of the artificial agent's ability to understand his emotional signals, he can decide to participate in the interaction exhibiting proper communicative signals.

Research problem and hypotheses

Combining human-machine interaction with a psychological view of emotion as a multiple component process (Scherer, 2000; Frijda, 1986; 2001; Roseman, 2001), this study focuses on the complex relationships existing among emotion, motivation and cognition in human-computer interaction, in particular:

1. Considering emotion as a process and its cognitive components (Lazarus, 1991; Scherer, 2001), we expect that:
   - tasks involving different cognitive appraisals will excite active responses on the environment that have an expressive component;
   - these responses consist of actions rather than reactions and are aimed at adapting to the environment and changing it (e.g. the subject puffs to signal the need to change the task);
   - during the task the subject will exhibit dynamic and flexible responses based on a continuous evaluative monitoring of stimuli rather than fixed expressive patterns corresponding to an emotional label;
   - each task will elicit congruent emotional expressive responses (e.g. boredom for the repetitive task).
2. A second research problem concerns the emotional interaction between the user and the machine. According to the Media Equation (Reeves, Nass, 1998) we suppose that during the interaction with the computer subjects will exhibit communicative signals (verbal and non verbal) to express their emotional state;
3. Finally, we'll try to answer to the following question: if the subject is aware of interacting with an understanding agent, is he encouraged to make use of emotional communicative signals? According to Fridlund (1991), we suppose that in the experimental condition (when the computer tells the subject that it is able to understand his/her emotional states) subjects will exhibit significantly more communicative signals than subjects in the control condition.

Method

Participants:

A total of 30 university students (20-23 years old) were recruited from two different kinds of faculties (humanistic vs. scientific).

Stimuli Construction:

Three different kinds of computer games were projected to modify the subject's attention level and to elicit specific emotional reactions. Systematic manipulation of appraisal dimensions was used through the selection of types of game events that were assumed to produce specific appraisals.

Specifically, game events were supposed to support four emotional evaluation checks: 1. novelty (a change in the situation that captures the subject's attention); 2. hedonic value (intrinsic pleasantness or unpleasantness of the stimulus); 3. goal conduciveness (events that help or damage the subject to reach a goal); 4. coping, (increasing levels of difficulty of tasks that change the estimated ability to deal with them). All games were previously tested on 10 subjects to assess their efficacy.

1. Quiz game: 15 questions are presented to the subject who has to select the right answer among four alternatives. The subject wins money for every correct answer and loses money when answering wrongly. Questions are divided into two series: a very easy one is followed by a very difficult one that makes the subject lose almost all the prize won. Selected events: correct/wrong answer; series of questions.

2. Boring game: the subject moves a rabbit on the screen and has to collect a great number of carrots (50). Carrots appear always in the same positions (repetitive task). A message appears every 10 carrots collected.

3. Enemy game: the subject moves a rabbit that has to collect carrots while avoiding an enemy. The game presents four different levels of difficulty. The subject wins points for every carrot collected and every level successfully completed. In each level positive or negative bonus appear randomly independently from the subject action. Selected events: losing life, passing to the next game level, positive/negative bonus.

Tools:

An enabling system was set up at the Psychology Communication Lab of the Catholic

University of Milan for implementing experimental sessions. Different kinds of devices were used to record the subject's behaviours (*see figure1*). All instruments were synchronized.

Figure 1: Experimental situation.

1. Two high resolution web cameras: one web-cam was placed in front of the subject to record facial movements, gaze direction and posture changes. A second camera was placed behind the subject so that it was possible for the external experimenter to follow the subject's action on the screen.
2. Physiological recordings were taken using the BIOPAC System (BIOPAC System, Inc.; Goleta, CA).
3. A high quality microphone to record vocal reports.

Procedure:

Subjects were asked to use the computer where an avatar (Baldi, CSLU Toolkit) guided them across the three different kinds of computer games. All sessions started with 2 minutes of free exploration of a web site for the baseline measure of physiological signals. Total duration was of about 20 minutes.

They were divided into two different groups, according to the kind of information received by the avatar. In particular, this experimental research will make use of the Wizard of Oz approach (Lisetti, 2001; Picard, Kort, Reilly, 2002), that is it employs an initially simulated prototype of emotional intelligent interface: the subjects are told they are interacting with an emotional intelligent computer, though in fact they are not.

In the **experimental condition**, the subjects were exposed to a simulated emotional-intelligent computer, where the avatar provided a simulated intelligent feedback to the user to decode his emotional state and to adapt the tasks accordingly. For example, the avatar used sentences like: «You seem to be bored, so here it is a new game» «You are in difficulties: I repeat the instructions for you». The simulated emotional-intelligent computer appeared to be automatic to the subjects, but it was actually controlled by an out of sight experimenter.

In the **control condition** the avatar guided the subjects in the same activities but did not simulate to decode emotion. All subjects were alone in the room with the computer.

At the end of the computer session, subjects were asked to answer to questionnaire and the Coping Inventory for Stressful Situations (Endler, Parker, 1990; Pedrabissi, Santinello, 1994). In the questionnaire subjects were asked: to assess their own abilities at using a computer; to judge on a 7 points rating scale the computer games according to emotional dimensions (surprising/ boring; pleasant/unpleasant; frustrating/enjoying); to judge the efficacy of the interaction with the computer. In particular they were asked to judge to which extent the computer had been able to understand their emotional states. In this way it was possible to test the efficacy of the simulated prototype, as 13/15 subjects in the experimental condition said they believed in the pc's ability to understand.

Data Analysis:

Different kinds of synchronized data were recorded:

Physiological signals:	ECG; Respiration; Galvanic Skin Response; Skin Temperature
Non verbal signals:	Posture; gaze direction facial movements
Non verbal vocal signals	supra-segmental cues
Verbal signals	speech

This is an in-progress study and data are still under elaboration. In this study we'll start to present the analysis of non-verbal communicative signals.

This choice is due to different reasons: 1. our hypothesis questions about the intentionality of non verbal signals; 2. non verbal signals are more continuous than verbal ones; 3. this analysis enables us to investigate the possibility of applying a second methodological level referring to dimensional axes.

Two different levels of analysis are going to be conducted. At a first level of analysis all video tapes are codified *frame by frame* (25 fps) using The Observer 5.0 NOLDUS Software and THEME Software for the recurrent pattern analysis. This analysis has previously required the elaboration of a coding grid and thus the selection of behavioural units to be extracted. Four macro-categories were then considered:

Facial movements: the fundamental muscle movements that comprise Facial Action Coding System (Ekman, Friesen, 1978; 2002)

were selected. We considered action units relating to upper face and lower face (20 AU and 10 AD). For each unit intensity was rated (Low, Medium, High).

Gaze direction: we consider if the subject looks at the screen, at the keyboard, around, etc.

Posture: behavioural units of moving near to /far from the screen were considered.

Vocal behaviours: in the video coding it was recorded when the subject speaks (verbal) or uses other kind of vocalizations (grumbling, no-words, etc). All vocal reports were even analysed through the Computerized Speech Laboratory KAY (CSL) Software for supra-segmental characteristics (pitch, volume, amplitude, time) analysis.

All behaviours (dependent variables) are analysed in relation to the following independent variables: 2 (experimental group) (between subjects) X 3 (kind of computer task) (within). Within each task the computer events are then considered.

At a second level of analysis, starting from the analysis of patterns of multimodal signals the mentioned dimensional axes (attention, hedonic value, coping) are scored on a 5 point rating scale from -2 to +2 by a group of judges trying to describe the user emotional state. Inter-judge agreement will be calculated (*k Cohen*).

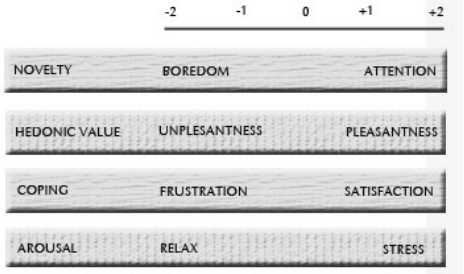


Figure 2: Dimensional axes.

Results:

This paper includes only initial results (corresponding to the analysis of one subject, 22 minutes and 27 seconds for a total of 26948 frames) which are indicative of the methodology that has been applied.

1. *Analysis of non verbal communicative signals*. Table 1 shows a first statistical output.

Results show that during the interaction with the computer, the subject exhibits communicative non verbal behaviour. In particular, facial behavioural units are used more frequently than vocal ones and within these facial movements there is a higher exhibition of lower face units (lips).

Table 1: (n) number of total occurrences; (tD) total duration (sec); (mD) mean duration (sec).

	n	tD	mD
Facial movements			
Inner brow raiser	67	20.04	0.30
Outer brow raiser	66	17.92	0.27
Brow lowerer	28	19.52	0.70
Upper lid raiser	26	13.16	0.51
Lid tightener	18	17.04	0.95
Chin raise	71	23.00	0.32
Cheek raise	36	61.32	1.70
Lip corner puller	63	102.16	1.62
Lip corner depressor	43	10.32	0.24
Lips part	99	523.68	5.29
Lip pucker	60	35.28	0.59
Lip pressure	78	52.64	0.67
Lips suck	14	6.68	0.48
Gaze direction			
Look at the screen	29	1296.32	44.70
Look at keyboard	17	10.28	0.60
Look around	8	21.44	2.68
Posture			
forward	22	1245.00	56.59
Head backward	15	87.92	5.86
Head forward	7	27.56	3.94
Initial position	5	41.24	8.25
backward	5	44.12	8.82
Vocal behaviors			
speech	29	31.88	1.10
laugh	7	5.96	0.85
Non words (oh, huh)	5	2.32	0.46
puff	2	2.00	1.00

The analysis of the most recurrent configurations of the emotional expressions show (Theme analysis: number of occurrences: 10; $p<.001$) the presence of patterns of different levels of complexity that involve facial movements, posture, etc. (figure 3).

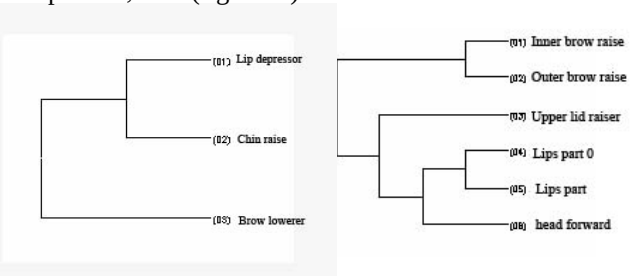


Figure 3: T-patterns of expressive behaviours.

Another interesting variable to examine concerns the time position of the emotional behavioural response in relation the antecedent action of the computer and the performance.

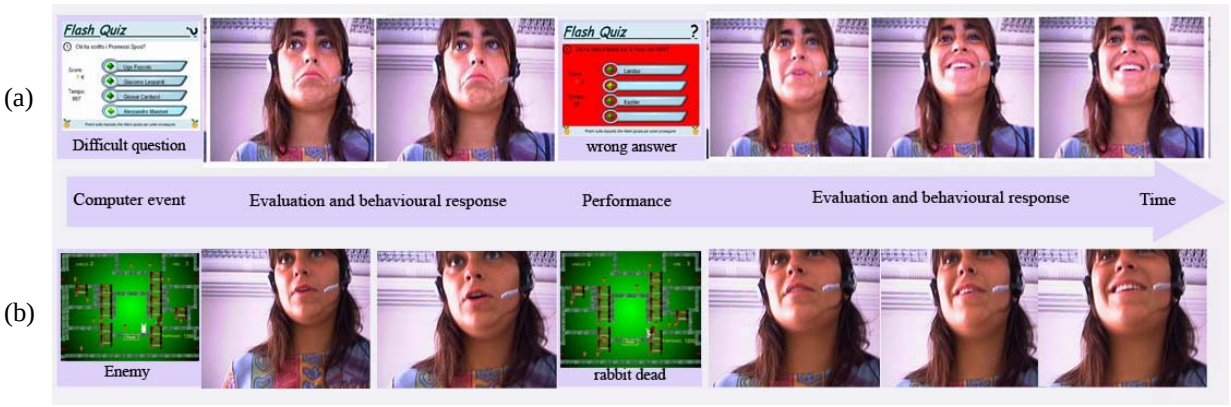


Fig. 4: Frame by frame sequences and time structure.

Figure 5 shows an example of two *frame by frame* sequences (6.89 sec; 5.32 sec) where the subject is dealing with the quiz game (a) and the enemy game (b). Each sequence starts with the computer event (difficult question; enemy): it is followed by its evaluation and a frustration behavioural response which precedes the subject's performance (wrong answer; rabbit dead). Finally the emotional response to the mistake (smile).

2. *Questionnaire data.* Finally Table 2 shows the answers of subjects about the interaction they experienced with the computer. Subjects in the experimental condition assign higher scores to the ability of the computer to understand emotion (a), to adapt the task according to them (b) and to answer to their reactions (c) while in the control condition subjects assigned high scores to the ability to give information (3).

Table 2. Mean values of the scores rated by subjects about the computer ability to understand.

Mean values	experimental	control
a	3,93	2,35
b	3,87	2,41
c	3,73	4,29
d	3,53	2,76

Discussion

As previously mentioned, this is a study in progress and data till now available does not allow us to discuss or support our hypotheses, in particular concerning the expected differences in the exhibition of communicative signals between the two conditions. Unfortunately this remains an open question to which we hope to answer in the following months. In any case, the results presented above enable some interesting considerations.

First of all, non verbal behavioural units are exhibited with different frequencies and durations, hence they seem to have different functions and relevance. It is possible that in the interaction with a machine subjects mostly rely on facial responses, even if vocal behaviour is present. Moreover, behavioural units are linked one to another in more complex pattern and expressive configurations. Hence the subject exhibits dynamical facial responses based on the continuous monitoring of the task and performance rather than fixed patterns corresponding to an emotional label. Secondly, patterns show a correlation between computer events and behavioural units: in this sense it has been possible to focus on sequences of events characterized by a specific timing and organization (antecedent, evaluation, performance, emotional response). In particular, the emotional response seems to have tight boundaries with the evaluation of the antecedent and the performance in the task suggesting its role in the motivation to act.

These considerations may have interesting applications when considering artificial agents and the implementing of computer interfaces as autonomous intelligences. In this perspective, knowing about what kind of emotional behaviours the human user is encouraged to show and direct to a supposed emotional-intelligent machine is a central issue to understand the kind of information that a computer can accurately obtain.

Secondly, the study presents an attempt to define some significant features that can be extracted from the user's expressive behaviour and a model able to describe the dynamic changing of the user's emotions during the computer interaction. This model can be seen as a possible formalization of an emotional semantics that could be used by a machine aimed at discerning the user's emotional state. As a matter of fact, within a multi-disciplinary attempt to implement an emotional intelligent interface (Andreoni, Apolloni, Balzarotti, Beverina, Ciceri, Colombo, Fumagalli, Palmas, Piccini, 2004) we suggest that such a machine should present not only physical devices

for capturing the user's multimodal signals, but also a specific component for features extraction and their semantic encoding.

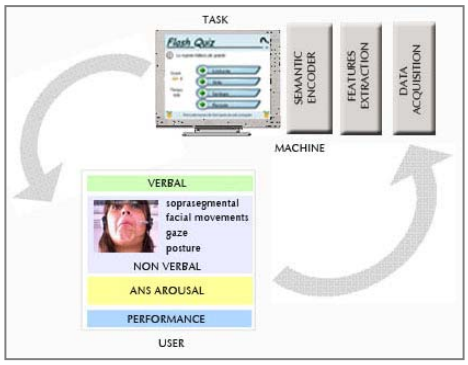


Figure 5: Theoretical model for an emotional intelligent interface

References:

G. Andreoni, B. Apolloni, S. Balzarotti, F. Beverina, R. Ciceri, P. Colombo, F. Fumagalli, G. Palmas and L. Piccini. Affective Learning. *Proceedings WIRN 2004 XV Italian Workshop on Neural Network, Perugia, 2004.*

Baldi CSLU ToolKit. Website http://cslu.cse.ogi.edu/toolkit/ .

L.W. Barsaou. Perceptual symbol systems. *Behavioral and Brain Sciences, 22:* 577-609, 1999.

C. Bartneck. How convincing id Mr Data's Smile: Affective expressions of Machines. *User Modelling and User-Adapted Interaction,* 11: 279-295, 2001.

Biopac systems inc (accessed mar 2004). http://www.biopac.com.

M.E. Bratman. What is intention? In: P.R. Cohen, J. Morgan and M.E. Pollack (Editors) *Intention in Communication,* Cambridge, MIT Press, 1990.

C. Breazeal. Sociable Machines: Expressive Social Exchange Between Humans and Robots. *Doctoral Dissertation*. Department of Electrical Engineering and Computer Science. MIT, 2000.

R. Ciceri. La paura. Il Mulino, Bologna, 2002.

R. Ciceri (Editor). *Comunicare il pensiero,* Omega Edizioni, Torino, 2001.

R. Ciceri and L. Anolli. La voce delle emozioni. Verso una semiosi della comunicazione vocale non-verbale delle emozioni. Franco Angeli, Milano, 1997.

R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor. Emotion recognition in human-computer interaction, *IEEE Signal Process*. 18: 32–80, 2001.

N. Bianchi, C. L. Lisetti. Modelling Multimodal Expression of User's Affective Subjective Experience. *User Modelling and User Adapted Interaction,* 12(1): 49-84, 2001.

P. Ekman, W.V. Friesen and J.Hager. Facial Action Coding System. The Manual. Research Nexus Division of Network Information Research Corporation, Salt Lake City, 2002.

N.S. Endler and J.D.A. Parker. Multi dimensional assessment of coping: concepts, issues and measurement. V European Conference on Personality, Ariccia, 1990.

A. Fridlund. Sociality of solitary smiling: Potentiation by an implicit audience. 1991. In: Parrott, W. Gerrod. *Emotions in social psychology: Essential readings.* New York, NY, US: Psychology Press. xiv, 265-280, 2001.

N. H. Frijda. The Emotions. New York: Cambridge University Press, 1986.

N. H. Frijda. The self and emotion. In: Bosma, A. Harke.; E. Kunnen and Saskia, *Identity and emotion: Development through self-organization. Studies in emotion and social interaction.* New York, NY, US: Cambridge University Press, xiv, 39-63, 2001.

J. Gratch and S. Marcella. Evaluating the modelling and use of emotion in virtual humans. *Proceedings of the 3rdInternational joint Conference on Autonomous Agents and Multi agent Systems,* 2004.

J. Gratch and S. Marcella. Tears and Fears: Modelling Emotions and Emotional Behaviours in Synthetic Agents. Presented at *AAMAS*, Montreal, Canada, 2001.

J.O. Greene (Editor). Message production: Advances in Communication Theory. Mahwah, N.J., Erlbaum, 1997.

A. Kappas, A. Pecchinenda. Don't wait the monsters to get you: A videogame task to manipulate to manipulate appraisals in real time. *Cognition and Emotion,* 13: 119-124, 1999.

A. Kappas, A. Pecchinenda. Rules of disengagement: Cardiovascular changes as a function of appraisal and nine levels of difficulty of an interactive video game task. *Psychophysiology* 37, S53, Abstract, 2000.

B. Kort, R. Reilly and R.W Picard. An Affective Model of Interplay Between Emotions and Learning: Reengineering Educational Pedagogy—Building a Learning Companion. *International Conference on Advanced Learning Technologies,* 2001.

G. Lakoff and M. Johnson. Philosophy in the flesh. The embodied mind and its challenge to western thought, New York, Basic Books, 1999.

R.S. Lazarus. Emotion and adaptation. New York, Oxford University Press, 1991.

C. L. Lisetti. Personality, Affect and Emotion Taxonomy for Socially Intelligent Agents. *In Proceedings of the 15th International Florida Artificial Intelligence Research Society Conference* (FLAIRS'02), Menlo Park, CA: AAAI Press, Pensacola, FL, 2002.

C. L. Lisetti, S.M. Brown, K. Alverez and A. H. Marpaung. A social informatics approach to human robot interaction with a service social robot. *IEEE Transactions on Systems, man and Cybernetics Special Issue on Human-Robot Interaction,* 2004.

C. L. Lisetti, F. Nasoz, C. LeRouge, O. Ozyer, and K. Alverez. Intelligent Affective Interfaces: A Patient-Modelling Assessment for Tele-Home Health Care. *International Journal of Human-Computer Studies*, 59: 245-255, 2003.

B. Mesquita, N.H. Frijda and K. R Scherer. Culture and emotion. In: J.W. Berry, P.R. Dasen, et al. *Handbook of cross-cultural psychology*, Vol. 2: *Basic processes and human development* (2nd ed.). Handbook of cross-cultural psychology. Needham Heights, MA, US: Allyn & Bacon. xxxvii, 255-297, 1997.

D. Norman. Emotional design: why we love (or hate) everyday things. New York, Basic Books, 2004.

B.J. O'Keefe and B.L. Lambert. Managing the flow of ideas: A local management approach to message design. In B.R. Burleson (Editor) *Communication Yearbook 18*, Thousand Oaks, CA: Sage, 1995.

L. Pedrabissi and M. Santinello. Coping Inventory for Stressful Situations: revision of validity and psychometric properties. *Ricerche di Psicologia*, 4 (18): 49-63, 1994.

R. W. Picard. Affective Computing. MIT Press. Cambridge, 1997.

R. W. Picard. What does it mean for a computer to "have" emotions? In: R. Trappl, P. Petta and S. Payr, *Emotions in Humans and Artifacts*, MIT Press, 2003.

R. W. Picard and J. Klein. Computers that Recognise and Respond to User Emotion: Theoretical and Practical Implications. *Interacting with Computers*, 14, (2): 141-169, 2002.

R. W. Picard, E. Vyzas, and J. Healey. Toward Machine Emotional Intelligence: Analysis of Affective Physiological State. *IEEE Transactions Pattern Analysis and Machine Intelligence*, Vol. 23 (10), 2001.

H. Prendinger, S. Mayer, J. Mori and M. Ishizuka. Using bio-signals to measure and reflect the impact of character-based interfaces. *Proceedings Fourth International Working Conference on Intelligent Virtual Agents* (IVA-03), 2003.

B. Reeves and C. Nass. The Media Equation. How People Treat Computers, Television and New Media Like Real People and Places. CSLI Publications, Centre for the Study of Language and Information. Cambridge University Press, 1998.

I.J. Roseman. Appraisal determinants of discrete emotions. *Cognition and Emotion*, 5, 161-200, 1991.

I. J. Roseman, A. Evdokas. Appraisals cause experienced emotions: Experimental evidence. *Cognition & Emotion*, 18(1): 1-28, 2004.

K.R. Scherer. Studying the emotion-antecedent appraisal process: An expert system approach. *Cognition and Emotion* 7:325-355, 1993.

K.R. Scherer. Emotions as episodes of subsystems synchronization driven by nonlinear appraisal processes. In M D. Lewis, I. Granic, (Editors) *Emotion, development, and self-organization: Dynamic systems approaches to emotional development.* Cambridge studies in social and emotional development New York, NY, US: Cambridge University Press. xiii, 70-99, 2000.

K.R. Scherer. Appraisal considered as a process of multilevel sequential checking. In: K. Scherer, A. Schorr, T. Johnstone, (Editors) *Appraisal processes in emotion: Theory, methods, research. Series in affective science*, London, Oxford University Press, xiv, 478 pp, 2001.

K.R. Scherer, A. Schorr and T. Johnstone. (Editors). Appraisal processes in emotion: Theory, methods, research. Series in affective science. London, Oxford University Press, 2001.

J.R. Searle. Mind, Language and society. New York, NY, US, Basic Books, 1998.

C. M. Van Reekum, T. Johnstone, R. Banse, A. Etter, T. Wehrle, and K.R. Scherer. Psycho physiological responses to appraisal dimensions in a computer game. *Cognition Emotion*, 18(5): 663-688, Aug 2004.

T. Wehrle and K.R. Scherer. Potential pitfalls in computational modelling of appraisal processes: A reply to Chwelos and Oatley, *Cognition and Emotion*, 9: 599-616, 1995.

B. Weiner. An attributional theory of motivation and emotion. New York, Springer, 1986.

Motivation Driven Learning of Action Affordances

Ignasi Cos-Aguilera*†, Lola Cañamero† and Gillian M. Hayes*

*IPAB, School of Informatics, University of Edinburgh,
Mayfield Road, JCMB 1.1106, EH9 1JZ Edinburgh, Scotland, UK.
ignasi@dai.ed.ac.uk, gmh@inf.ed.ac.uk

†School of Computer Science, University of Hertfordshire,
College Lane, Hatfield, Herts, AL10 9AB, UK.
L.Canamero@herts.ac.uk

Abstract

Survival in the animal realm often depends on the ability to elucidate the potentialities for action offered by every situation. This paper argues that affordance learning is a powerful ability for adaptive, embodied, situated agents, and presents a motivation-driven method for their learning. The method proposed considers the agent and its environment as a single unit, thus intrinsically relating agent's interactions to fluctuations of the agent's internal motivation. Being that the motivational state is an expression of the agent's physiology, the existing causality of interactions and their effect on the motivational state is exploited as a principle to learn object affordances. The hypothesis is tested in a Webots 4.0 simulator with a Khepera robot.

1 Introduction

One of the most vital abilities for situated, embedded, autonomous agents in a dynamic scenario is making the right decisions when interacting with their environment. This is the so-called *behaviour or action selection* problem, deciding "what to do next" (what behaviour to execute in a particular situation) to increase the likelihood of maintaining life. Being able to make the right decisions partly depends on the knowledge of the effect of an action to compensate internal needs. Furthermore, it depends on the ability to discriminate objects to befit every interaction. This was confirmed experimentally by Guazzelli et al. (1998), who proposed a behaviour selection model to simulate the behaviour of rats navigating a T-maze, integrating drives and affordances. No perception-related learning was however involved, being that this was solely aimed at interpreting the possibility of moving in one or another direction.

The use of motivational states to make decisions has been proposed in several architectures (Avila-García and Cañamero, 2002; Cañamero, 1997), which mention the necessity not only of maintaining life, but also of meeting the criterion of internal physiological stability (Ashby, 1965). Nevertheless, these architectures neglect the apprehension of the appropriate functionalities of objects. Information about the objects' potential for action has therefore usually been hard-wired. It is argued that knowing the functionality of an object is also part of the adaptation problem.

Related to this, Gibson introduced the notion of *affordance* (Gibson, 1966), defined as the functionality an object offers to an agent. Hence, a set of affordances is only defined in the context of a particular agent-environment pair. Furthermore, *affordances are held to be directly available from the environment, without the integration of perceived features into object representations* (Cooper and Glasspool, 2002). Based on this, Cooper and Glasspool (2002) introduced a symbolic model of affordance learning by relating object features to action schemas. In their approach, object features are symbolically integrated into objects to bias one action or another.

Conversely, the architecture introduced in this paper aims at endowing the agent with the capability of building its own functional perception via an appropriate neural representation of the objects in its environment, related to the agent's behaviour repertoire[1]. This aims at bypassing the feature-based step, and should therefore be a more faithful implementation of gibsonian affordances. Importantly, to perform an action the perception of certain regularities of each object is fundamental to decide the right be-

[1]Unlike Gibson's studies of the optical flow, we have to deal with other perceptual modalities (the agent's senses).

haviour. However, this does not relate to physical resemblance only (among different objects of the same sort), but also to a functional similarity (being able to perform the same actions).

The next section introduces the affordance learning and behaviour selection model, and precedes the experimental section. The paper concludes with a discussion of results and of future research issues.

2 Motivational Model for Learning Affordances

The model comprises three parts: a neural structure, a behaviour arbitration mechanism and a learning module.

2.1 Neural Structure

The first challenge is to build a neural representation of the objects of the environment. To this end, the use of a Growing When Required (GWR) network (Marsland et al., 2002) has been selected. This is a topological network that adapts to the level of entropy of the environment according to a set of parameters, unlike Kohonen (1982). The growing process is described in the following steps:

1. The network is trained with 64-D image patterns representing objects in the scenario. The algorithm chooses the first and second most similar nodes.

2. If the Euclidean distance between the closest node and the current interaction pattern is larger than the pre-set accuracy, a new node is inserted between the two closest nodes, which are then connected by new synapses. Conversely, the closest and its adjacent nodes are dragged towards the input pattern.

3. Nodes rarely close to the patterns are deleted.

4. The growing process is hindered when the euclidean distance between the sensory-patterns and their closest node is smaller than the pre-set level of accuracy.

In a very simple manner, the GWR provides a simple representation of similar objects. The next subsection explains how to relate these patterns to the behaviour repertoire.

2.2 Motivations for Behaviour Selection

The combination of extenal and internal stimuli gives rise to the motivational state. This section describes the necessary elements to build an internal physiology.

The controlled homeostatic variables are abstractions representing an agents' resources. Nutrition, stamina and restlessness are the chosen variables. Their values must be kept within the *viability zone* for the agent to remain alive; if their values overflow/underflow the upper/lower boundaries that define the variable's viability, the robot dies.

The drives are also abstractions denoting urges for action. The drives monitor the levels of the homeostatic variables and initiate a process of compensation whenever they are in a deficit state. In our case, the mechanism of compensation is the selection and execution of a behaviour, which requires an appropriate object nearby for successful execution. We have used three different drives: hunger (which controls nutrition), fatigue (controlling stamina), and curiosity (controlling restlessness). At each time step, each drive is assigned an intensity proportional to the magnitude of the error of its controlled variable.

The behaviours are to grasp, to shelter and to interact. The execution of a behaviour results in an interaction with an object in the environment that may cause a compensation of the deficit for the most critical internal variable, contributing therefore to compensate the drives. In the general case, different behaviours can contribute to compensate a drive, but in our simplified model each drive can be satisfied by one behaviour only, "eat" (grasping an object) satisfies hunger, "shelter" satisfies fatigue, and "interact" satisfies curiosity.

The arbitration mechanism for behaviour selection follows a winner-take-all policy, using the drive that exhibits the highest urgency (the one with the highest level) to choose the behaviour to execute next. In our simplified model this is very straighforward because there is a single behaviour that can satisfy each drive.

The model also has two *Hormones*: Frustration and Satisfaction, which are respectively triggered when the outcome of an interaction succeeds or fails. The values of the hormones are 1, if they are active, and 0 otherwise.

2.3 The Learning Mechanism

The learning process adds a novel dimension to the topological network, by growing *functional* synapses between every node in the aforementioned neural structure and each behaviour of the agent. The pro-

cedure for growing these synapses is driven by the agent's drives in a hebbian manner. The process is as follows:

- Every time the agent detects an object, the closest node in the state space is identified. Figure 1 shows the 2D projections of topologies representing the objects contained in the Khepera world used for simulation.
- The interaction succeeds, the hormone Satisfaction is released, otherwise, the hormone Frustration is released.
- Satisfaction and Frustration, strengthen or weaken, respectively, the synapse relating the active node and the behaviour executed ($\Delta\omega_{ij} = \alpha b_j$). Weights are normalised between -1.0 and 1.0.

The final values quantify the *affordances* relating those particular objects, encoded by the neural structure, to the agent's behaviours.

3 Experiments and Results

The goal of these experiments is to test this learning hypothesis with an artificial agent in an engineered scenario. The affordances of the objects in that scenario are such that little objects afford grasping, large objects afford shelter, and all objects afford interacting. Relative sizes vary between 0.08 and 0.01, the size of the Kephera's gripper is 0.04 and the arena measures 0.5×0.5 units.

3.1 Experimental Procedure

The robot wanders in the aforementioned environment, interacting with objects encountered at random. Everytime an object is encountered, the object is centred, and a snapshot of the object is taken always at the same distance. The single top horizontal line of the object is selected, and reduced to a 64-D illumination vector. This vector is used for building the neural structure[2]. Two 2D-PCA of final structures with 16 and 42 nodes each are shown in figure 1.

Concurrently, the agent's homeostatic variables are initialised to their optimal value, and decay according to equation $\Delta hv_i = \tau$, with $\tau = 10^{-5}$. Their optimal values are 0.8 for nutrition and stamina, and 0.2 for restlessness. Their related drives measure the difference from those optimal values, and define the agent's *motivational state*. Whenever an object is encountered, the behaviour whose attached drive exhibits the highest value is selected and executed. The *affordance learning* method, as introduced in section 2.3, is then executed.

[2]With parameters $energy = 0.5$, $epsilon_b = 0.5$, $epsilon_n = 0.006$, a_{max}=50, as described in Marsland et al. (2002).

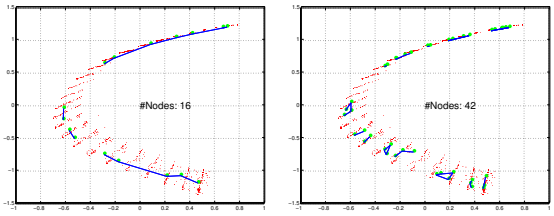


Figure 1: 2D-PCA with GWR overlapping with 16 and 42 nodes, left and right, respectively.

3.2 Results

Four series of five simulations each have been run with networks of sizes between 4 and 42 nodes for testing the aformentined learning algorithm. Results for topological networks of 4, 8, 16 and 42 nodes are presented in histograms 2 and 3. The three individual histograms, address the affordance values for each behaviour: grasp, shelter and interact. Values in the X-axis represent the node id in the topological structure, and values in the Y-axis the affordance values learnt (ranged between -1.0 and 1.0), averaged over five simulations.

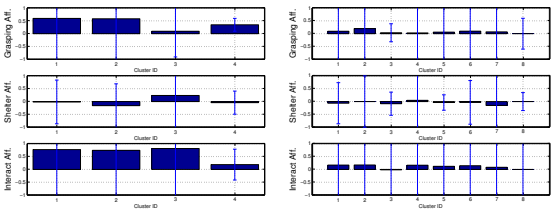


Figure 2: Learnt affordance values for behaviours grasp, shelter and touch (top-down) for GWR with 4 and 8 nodes, left and right, respectively.

It can be observed that affordance values in topologies with a low number of nodes exhibit a large standard deviation. This is due to the low accuracy of those topologies, and is confirmed by observing the difference with affordance values in topologies with a larger number of nodes (16 and 42), which are defined more precisely. In the former case, the low level of accuracy provokes an incorrect selection of the node closest to the visual pattern. In other words,

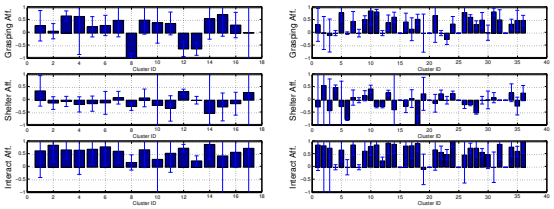


Figure 3: Learnt affordance values for behaviours grasp, shelter and touch (top-down) for GWR with 16 and 32 nodes, left and right, respectively.

nodes in topologies with a low level of accuracy represent a range of objects whose features cannot be causally related to the same effect. This also highlights that for representations of high accuracy, the growing algorithm could be improved via pruning nodes exhibiting affordance values with a high variance. This would not diminish the overall performance, since the resting nodes already represent the sensory-space accurately enough. This would improve the overall performance, since the selection of one node or another would be more accurate, thus its affordance values would be better defined.

Lastly, it is important to highlight that there are implementation and execution issues, e.g., inaccurate object manipulation, which means the execution of some behaviours fail despite the object affording that behaviour to be executed.

4 Conclusions and Future Work

The learning method is based on internal observation of causal fluctuations in the motivational state due to behaviour execution. This provokes a hormonal response, which reinforces the functional synapses relating the behaviour executed to the node in the GWR closest to the perceived sensory pattern. The results suggest that affordances can be learnt according to the experimental procedure proposed.

It is fundamental to stress that affordances are context-related. Hence, to be able *to learn and use affordances, it is necessary to define a context*: the agent's morphology, its set of internal goals and behaviours, the environment. However, sensory perception is independent from the motivational state.

The principles of the model highlight that *motivation and learning are two inter-related processes*. If there is motivation to drive the agent to perform an action, the effect of the performance biases learning. Conversely, learning has a reinforcing role on the motivational (physiological) system. This is grounded in neuroscience by Bindra's suggestion: *"The effects on behaviour produced by reinforcement and motivation arise from a common set of neuro-psychological mechanisms, and the principle of reinforcement is a special case of the more fundamental principle of motivation"* (Bindra, 1969).

Finally, it is relevant to stress that *learning affordances is related to building a representation of the environment; however, a functional representation*. In fact, as the model shows, neural encoding and reinforcement are processes affecting one another.

Future endeavours will perform ethological analysis of behaviour (in terms of physiological stability and cycles of behaviour execution), to assess the effect and reach of this learning process in a variety of environments.

References

W.R. Ashby. *Design for a Brain: The Origin of Adaptive Behaviour*. Chapman & Hall, London, 1965.

Orlando Avila-García and Lola Cañamero. A comparison of behaviour selection architectures using viability indicators. In *Proc. of International Workshop on Biologically-Inspired Robotics: The Legacy of W. Grey Walter*. Bristol HP Labs, UK., August 2002.

D. Bindra. The interrelated mechanisms of reinforcement and motivation, and the nature of their influence on response. In W. J. Arnold and D. Levine, editors, *Nebraska Symposium on Motivation*, pages 1–33. University of Nebraska Press, 1969.

Lola D. Cañamero. Modeling motivations and emotions as a basis for intelligent behavior. In W. Lewis Johnson, editor, *Proceedings of the First International Symposium on Autonomous Agents (Agents'97)*, pages 148–155. New York, NY: ACM Press, 1997.

R. Cooper and D. Glasspool. Learning affordances and action schemas. In R.M. French and J. Sougne, editors, *Connectionist Models of Learning, Development and Evolution*, pages 133–142. Springer-Verlag: London, 2002.

James J. Gibson. *The Senses Considered as Perceptual Systems*. Houghton Mifflin Company, Boston, 1966.

A. Guazzelli, F. J. Corbacho, Bota M., and M.A. Arbib. Affordances, motivation, and the world graph theory. *Adaptive Behavior*, (6(3/4)):435–471, 1998.

Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69, 1982.

Stephen Marsland, Jonathan Shapiro, and Ulrich Nehmzow. A self-organising neural network that grows when required. *Neural Networks*, 15(8-9):1041–1058, 2002.

Figurative language expressing emotion and motivation in a web based learning environment

Manuela Delfino[*†]
[*]Institute for Educational Technology, CNR
Via de Marini 6 – 16149 Genova (Italy)
delfino@itd.cnr.it

Stefania Manca[†]
[†]Institute for Educational Technology, CNR
Via de Marini 6 – 16149 Genova (Italy)
manca@itd.cnr.it

Abstract

The present study aims to investigate how figurative language was used by the participants of an online learning environment in order to express their emotion, feelings and motivation in their new learning experience. According to our results, figurative language mainly served as an original and specific linguistic feature through which people project themselves (their identity, emotions and feelings) in the online context. The research was conducted on a ten-week course, delivered at a distance via a computer conferencing system, addressed to 57 student teachers.

1 Introduction

The study was conducted on the premises that cognitive processes are closely related to the affective, emotional and motivational ones. This kind of strict relationship is borne out within the context of neurosciences (Damasio, 1994; LeDoux, 1996), Artificial Intelligence (Picard, 1997; Dautenhahn, Bond, Cañamero and Edmonds, 2002), cognitive psychology (Hamilton, Bower and Frijda, 1988; Oatley and Jenkins, 1996; Frijda, Manstead and Bem, 2000) and social sciences (Elster, 1999; MacMullen, 2004), with positive effects on the educational context (Gardner, 1983; Volet and Järvelä, 2001) and on the context of web-based learning as well.

In the latter, the socio-affective dimension of learning acquires specific features as it is expressed and biased by written discourse. While some early approaches in the study of CMC emphasize the lack of non-verbal cues as a limit (Short, Williams and Christie, 1976), in more recent times, a number of studies have shown that written communication is able to actively stimulate and effectively enhance social and affective presence (Garrison, Anderson and Archer, 2000).

In contrast with the view according to which students adopt verbal immediacy behaviours to make up for the lack of non verbal and vocal cues communicated online, we reckon that the expression of emotions and affectivity must not be seen as a substitute or a surrogate way to express the same emotional needs that may emerge in a face-to-face setting, but rather as a different and independent means to become aware of and to share emotional states. In this view, written communication is supposed to convey specific and unique social and emotional affordances (Kreijns, Kirschner and Jochems, 2002).

Following this line of thought, we investigated the uses of figurative language as one of the possible dimensions adopted to express emotions in online communication, quite often achieved by using language in original and creative ways. Hence our analysis covered not only metaphors and other figures of speech in their proper meaning, but any use of language aiming at expressing a non literal meaning, i.e. a meaning beyond the standard denotation of the used utterances.

2 Theoretical background

The study of emotions in online learning has been carried out through a number of indicators: the main emotions involved in the experience of starting a distance education course (Conrad, 2002; O'Regan, 2003), and student distress in a web-based course (Hara and Kling, 2000). McFadden, Herie, Maiter and Dumbrill (in press) propose a model of web-based education based on the assumption that emotional emphasis may facilitate constructivist learning goals.

The role of students' online emotional appraisal of social conditions of learning has been studied in connection with research on motivation as well: emotion arousal influences the cognitive, metacognitive and motivational aspects of the

learning process, especially when they are socially oriented (Wosnitza and Volet, in press).

The affective and emotional functions of metaphors have been closely investigated in a number of studies as well. According to Lakoff and Johnson's theory of metaphor (2003), emotion concepts emerge as conceptual structures largely constituted by metaphor: emotion concepts are claimed to be social-cognitive constructions (Kövecses, 2002). Ortony and Fainsilber (1989) underlined concrete vividness as the main characteristic of metaphor and figurative language in the expression of emotions.

More generally, some authors (Gibbs, Leggitt and Turner, 2002) stated that figurative language is so special as it concerns emotional communication, which tightly reflects something about people's ordinary conceptualizations of their complex emotional experience. In addition, it is considered a special communicative tool because it might create that sense of closeness and intimacy between speaker and listener that literal language cannot achieve (Fussell and Moss, 1998).

3 The method

The research context was a ten-week course delivered at a distance via a computer conferencing system, held during the academic year 2002/2003 by our institute, on the topic of educational technology. The course was addressed to 57 student teachers of the local Post-graduate School for Secondary School-teachers and was managed by 7 tutors.

After the course conclusion we noticed the great amount of figurative language produced by tutors and students that occurred in their written discourse as a spontaneous phenomenon.

Next to the instructional activities, the use of some familiarization and meta-reflection facilities was especially encouraged for socialization and reflection purposes. Focusing the analysis on expressions of self-disclosure, two communication areas (socialization and meta-reflection) were object of investigation, being the most concerned with the expression of emotions.

Following the most used approach in the literature about computer-mediated discourse analysis (i.e. Rourke, Anderson, Garrison and Archer, 2001; Herring, 2004), the single message (*posting*) was chosen as macro-unit of analysis, since it was recognized as the smallest meaningful, independent and exhaustive datum. Postings were analyzed to find the cases in which figurative language served to express the participants' feelings. As a single posting could include more than a single figurative language instance (*occurrence*), segments of postings were considered for both quantitative and qualitative micro-analysis. The research was guided by the following questions:

- which images did the participants choose to introduce themselves to others?
- which images did they use to represent their learning experience?
- what kind of emotions were shared using figurative language?

4 Qualitative outcomes

The number of postings with figurative language was equivalent to 86 of 843 examined (10.2%), and the number of occurrences was 103, postings containing on average 1.19 occurrences.

During the analysis it was noticed that all the occurrences could belong to two alternative categories: some were related to the expression of participants' Identity, some others to the expression of feelings and ideas towards the Context of the course (see Table 1).

Table 1: Categories chosen in order to analyze figurative language occurrences

Categories	Category's typology	Iconic image
Identity	I feel like…	disguise
	We feel like…	
	I see you as…	
	I see them as…	
	I move as…	orientation
Context	CMC environment is…	give body
	Written communication is…	
	The course is…	
	The course subject is…	
	The computer is…	give soul

Participants used figurative language both to give shape and body to themselves and to other participants, disguising their corporeity and making it move in different settings, and giving a body and a soul to objects. In other words, they recurred to figurative language with the effect of changing the shared ontological status of people and objects.

Thinking of the reasons why participants chose to use figurative language, emerged quite clearly that by acting as other people, or dressing up as animals, literature characters, cars and so on, they could explain their inner emotions, such as fear, frustration, anger, happiness, moderating, at the same time, their epistemic commitment towards the propositional content of their statement. For instance, for some participants it might have been difficult to explicitly acknowledge that they were very anxious because they could not understand what was going on in the online course, but they had

no problem to state that they were in need of a lifeboat, since they felt quite shipwrecked (see below the full excerpt). Such disclosure was possible only with the reduced degree of epistemic commitment granted by figurative language.

4.1 Figurative language occurrences of Identity and Context

Especially during the first weeks of the course, participants recurred to the semantic field of navigation in order to express feelings related to the new learning environment.

In this way a little boat became, to all purposes, a vehicle useful both to represent a route and to explain feelings towards the learning experience:

"Until now my little boat passed off, without too many hitches…" - (II week)

This figurative idea was further developed by another student:

"Picking up the metaphor used by Irene, during this online activity I feel as I am in a paradoxical condition, on one side I'm navigating on the paper-boat of my "empiric" and improvised ICT competences, on the other, however, it seems to me that I'm sailing safe in this environment" - (II week)

Unfortunately, not all the seas sailed by the participants are so calm:

"Yesterday I'd like to use a virtual lifeboat; I felt as a shipwrecked" - (II week)

But not all the settings of figurative occurrences are placed in the sea. In order to explain the feelings towards their rhythm of participation to the activities, somebody wrote:

"In this brand-new activity, I feel somehow as I was a little turtle going slowly, slowly, slowly…" - (IV week)

and another participant echoed:

"I'm going slowly, uphill, but as an old 500[1] *I'm proceeding determined, one step at a time, always trying to learn something new and astonishing"* - (VII week)

The computer is the communication medium in the CMC environment and however transparent it may be, some participants felt its presence. In a posting, a student is invited to make a propitiatory gesture by ignoring the computer presence:

"You are MAD! Don't you know that these devices have eyes, ears and tongue? Don't you know that they love teasing and feel at the centre of attention? Of course I'm joking: PC infected me!" - (II week)

The course subject, as well as some postings and the reflection around synchronous and asynchronous communication, are objects of simile or comparison. A single posting might be a symbol of hope:

"I'm very grateful to the latter posting written by Giovanna. In this world, full of anxieties, a reassuring posting is like a dewdrop in the desert. Thank you" - (II week)

And again:

I'd like to thank dear and nice Irene for her appraisal to my posting: you don't know how much I appreciate that you sense a smell of "life" in my message - (IX week).

5 Conclusions and future directions

Self-disclosure has been an effective mean through which people reciprocally invite to open the door to motivation through the expression of emotions and feelings. All the examples reported a set of emotions of self-, other- and technology-directed type, expressed by means of figurative language. This latter has been a powerful detector of emotions and feelings involved in the learning experience (for a detailed analysis, Delfino and Manca, u.c.).

Figurative language has been a resource, among others, to create the new learning and social reality in which the participants were involved. For most of them it was their first online learning experience and they had to face several new problems including learning to communicate by written discourse in an asynchronous manner, familiarizing with communication technologies, and even practicing with learning and collaborating in group. Metaphors and figurative language helped them to understand a new domain of experience in terms of what was already familiar to them (e.g., images of movement to explain learning rhythms).

In the future, figurative language could also be adopted in the design and conduction phases of an online learning course, as a stimulus and a motivation to manifest and share those personal emotions and feelings always deeply involved in any new learning experience, by providing a framework for role assignment, identity, responsibility and intrinsic motivation.

References

D. Conrad. Engagement, excitement, anxiety and fear: learners' experiences of starting an online course. *The American Journal of Distance Education*, 16(4):205-226, 2002.

A. R. Damasio. *Descartes' Error: Emotion, Reason, and the Human Brain*. Grosset/Putnam, New York, 1994.

[1] A FIAT 500 is a popular Italian car from the 60's, whose most distinguishing features are the compact dimensions combined with toughness and reliability.

K. Dautenhahn, A. Bond, L. D. Cañamero and B. Edmonds (eds.). *Socially Intelligent Agents: Creating Relationships with Computers and Robots*. Kluwer Academic Publishers, Norwell, MA, 2002.

M. Delfino and S. Manca. The expression of social presence through the use of figurative language in a web-based learning environment. *Computers in Human Behavior,* under consideration.

J. Elster. *Alchemies of the Mind: Rationality of the Emotions*. Cambridge University Press, Cambridge, UK, 1999.

N. H. Frijda, A. S. R. Manstead and S. Bem. (eds.). *Emotions and Beliefs. How Feelings Influence Thoughts*. Cambridge University Press, Cambridge, UK, 2000.

S. R. Fussell and M. M. Moss. Figurative language in descriptions of emotional states. In S. R. Fussell and R. J. Kreuz. *Social and cognitive approaches to interpersonal communication*. Lawrence Erlbaum Associates, Mahwah, NJ, 1998.

H. Gardner. *Frames of Mind. The Theory of Multiple Intelligences*. Basic Books, New York, 1983.

D. R. Garrison, T. Anderson and W. Archer. Critical inquiry in a text-based environment: computer conferencing in higher education. *The Internet and Higher Education*, 2(2/3):87-105, 2000.

R. W. Gibbs, J. S. Leggit and E. A. Turner. What's Special About Figurative Language in Emotional Communication?. In S. R. Fussell. *The Verbal Communication of Emotions. Interdisciplinary Perspectives*:125-149, Lawrence Erlbaum Associates, Mahwah, NJ, 2002.

V. Hamilton, G. H. Bower and N. C. Frijda (eds). *Cognitive Perspectives on Emotion and Motivation*, Kluwer Academic Publishers, Boston, MA, 1988.

N. Hara and R. Kling. Student distress in a web-based distance education course. *Information, Communication & Society*, 3(4):557-579, 2000.

S. C. Herring. Computer-Mediated Discourse Analysis. An Approach to Researching Online Behavior. In S. A. Barab, R. Kling and J. H. Gray. *Designing for Virtual Communities in the Service of Learning*: 338-376, Cambridge University Press, Cambridge, UK, 2004.

Z. Kövecses. Emotion Concepts: Social Constructionism and Cognitive Linguistics. In S. R. Fussell. *The Verbal Communication of Emotions. Interdisciplinary Perspectives*:109-124), Lawrence Erlbaum Associates, Mahwah, NJ, 2002.

K. Kreijns, P. A. Kirschner and W. Jochems. The sociability of computer-supported collaborative learning environments. *Journal of Education Technology & Society*, 5(1):8–22, 2002.

G. Lakoff and M. Johnson. *Metaphors we live by*. Chicago University Press, Chicago, 2003.

J. LeDoux. *The emotional brain. The mysterious underpinnings of emotional life*. Simon & Schuster, New York, 1996.

R. J. MacFadden, M. Herie, S. Maiter and G. C. Dumbrill. Achieving high touch in high tech: A constructivist, emotionally-oriented model of web-based instruction. *Journal of Teaching in Social Work*, in press.

R. MacMullen. Historians Take Note: Motivation = Emotion. *Diogenes*, 51(3):19-25, 2004.

K. Oatley and J. M. Jenkins. *Understanding emotions*, Blackwell, Oxford, 1996.

K. O'Regan. Emotion and E-learning. *Journal of Asynchronous Learning Networks*, 7(3), 2003.

A. Ortony and L. Fainsilber. The role of metaphors in descriptions of emotions. In Y. Wilks. *Theoretical issues in natural language processing*: 178-182, Lawrence Erlbaum Associates, Mahwah, NJ, 1989.

R. W. Picard. *Affective computing*. The MIT Press, Cambridge, MA, 1997.

L. Rourke, T. Anderson, D. R. Garrison and W. Archer. Methodological Issues in the Content Analysis of Computer Conference Transcripts. *International Journal of Artificial Intelligence in Education*, 12:8-22, 2001.

J. Short, E. Williams and B. Christie. *The social psychology of telecommunications*. Wiley, London, UK, 1976.

S. E. Volet and S. Järvelä (eds.). *Motivation in learning contexts: Theoretical advances and methodological implications*, Elsevier Science, Amsterdam, 2001.

M. Wosnitza and S. E. Volet. Significance of social and emotional dimensions in online learning. *Learning and Instruction*, in press.

EXPERIMENTAL STUDY OF EMOTIONAL MANIFESTATIONS DURING A PROBLEM-SOLVING ACTIVITY

Delphine Duvallet*
Université de Rouen
Laboratoire de Psychologie et Neurosciences de la Cognition
Psy.Co (EA 1780)
76821 Mont Saint Aignan Cedex – France
*duvallet_delphine@yahoo.fr

Evelyne Clément**
Université de Rouen
Laboratoire de Psychologie et Neurosciences de la Cognition
Psy.Co (EA 1780)
76821 Mont Saint Aignan Cedex – France
**evelyne.clement@univ-rouen.fr

Abstract

The aim of the present research was first to identify the manifestations of emotion during a problem-solving activity and then to integrate, in a foreseeable future, the emotional component of the cognitive activity into the Constraints Model, a problem-solving model developed by Richard (Richard, Poitrenaud & Tijus, 1993; Richard & Zamani, 2003). Nineteen female participants were asked to solve the well-known five-disk version of the Tower of Hanoi problem. Two classical measures of emotion were recorded during the problem-solving activity: facial expressions and spontaneous skin conductance activity. The analysis of data shows that both facial and physiological manifestations do not randomly appear in the course of the problem-solving but in some cognitive significant events like the impasse situations in which the subject have to cope with actions sequence that are not relevant in sub-goal achievement. Those results suggest that emotion plays a crucial role in decision-making mechanism and that one of its functional aspects is to regulate the system.

1 Introduction

Since four decades, some cognitive investigators contribute to a highly important aspect of emotion theory by integrating it into a cognitive science framework. For instance, as far back as 1967, Simon claimed the necessity to develop a general theory of thinking and problem-solving that incorporates motivation and emotion. Simon (1967) and Oatley and Johnson-Laird (1987) considered that the function of emotion is to regulate the system by allowing to abandon the current goal and to substitute a new goal more suitable for the constraints of the environment.

Recently, there have been some attempts to model emotion within cognitive science framework (e.g. Gadanho & Hallam, 2002; Sander & Koenig, 2002; Gratch & Marsella, 2004).

For instance, Belavkin proposed a model based on the ACT-R architecture (Anderson & Lebière, 1998) that takes emotion into account (Belavkin, Ritter & Elliman, 1999; Belavkin, 2001) in a problem-solving activity. The author concludes that emotions make a positive influence on problem-solving by 1) increasing of the motivation and confidence when positive emotions are experienced on successes during problem-solving, 2) overcoming possible problems and allowing to change the direction of the search when negative emotions are experienced on failures during problem-solving.

Following Belavkin, we think that even in a situation carried out in a laboratory, problem-solving situations generate emotions that are experienced on successes or on failures to achieve the goal. According to the appraisal theories (e.g. Smith & Lazarus, 1990; Ellsworth, 1991; Lazarus, 1991; Roseman, 1991; Scherer, Schorr, Johnstone, 2001), in which the cognitive appraisal of the situation is an initial step in the triggering of emotion, achieving sub-goals should produce positive emotions, whereas leading to an impasse should produce negative emotions.

The aim of the present research was first to identify the manifestations of emotion during a problem-solving activity and then to integrate, in a foreseeable future, the emotional component of the cognitive activity into the Constraints Model, a Problem-Solving Model developed by Richard (Richard, Poitrenaud & Tijus, 1993; Richard & Zamani, 2003). This experimental step is a crucial one to integrate emotion in a cognitive model. In the Constraints Model, solving a problem consists to elaborate the adequate representation of the problem by dropping misconceptions (inappropriate

interpretations and irrelevant goals generated by these misconceptions) and building more and more sophisticated goal-structures (Richard, 1999).

On the one hand, this model is able to describe the solver behaviour into the two phases depicted by Kotovsky and Fallside (1989) in the course of problem-solving: the exploratory phase during which the relevant interpretation of the operator is elaborated, and the final phase in which planning takes place. On the other hand, the model contains a decision mechanism, which states that, in impasse situations, the representation temporally changes in order to make an action (principle of constraints-relaxation). In this approach, the impasse and the achievement of a sub-goal are supposed to be the occasion to change the representation (when the first representation lead to impasse) and the goal (both when the first representation lead to impasse, and when a sub-goal is achieved).

Our first hypothesis is that if the task is emotionally relevant to the subject, emotional manifestations may be more frequent during the exploratory phase than in the final phase. Our second hypothesis is that if the function of emotion is to regulate the system, so emotion may play a crucial role in decision-making mechanism and emotional manifestations may be observed during the significant events that are the impasses.

2 Method

Nineteen female participants were asked to solve the five-disk version of the Tower of Hanoi problem.

Two classical measures of emotion were recorded without interruption during the problem-solving activity: facial expressions and spontaneous skin conductance activity (Biopac's finger electrodes). Participants were tested individually and videotaped. As an indirect measure of emotion, facial expressions were chosen because participants had to solve the problem in the presence of the experimenter. Indeed facial expressiveness is generally considered as a component of emotion, of which the function is to communicate to someone his/her own internal states (e.g. Scherer, 1984; Frijda, 1986; Kaiser & Wehrle, 2001). In other respects, it seems important to measure the emotional manifestations as well through their physiological response since this response is not under the subject's control (Bauer, 1998). The analysis of the physiological data was based on two criteria: the average level of skin conductance, as reflecting the resources mobilisation (Dawson, Schell & Filion, 1990) and the maximum skin conductance amplitude (pic-to-pic), as reflecting the mobilisation's stability that is supposed to change in response to emotionally relevant events (Frijda, 1986). In order to identify for each subject in the course of the problem 1) the two phases described below and, 2) the impasses, we conducted a subject-by-subject analysis in the framework of the Constraint Model (Richard et al., 1993). This analysis allowed us to match the emotional manifestations to the significant events of the problem-solving activity.

3 Results

The three measures of emotion are compared according to the solving phases (exploratory versus planning), and within the exploratory phase, they are compared between impasse situations and the remaining exploratory phase.

3.1 Exploratory phase versus planning phase

The average number of facial expressions per minute (table 1) is more frequent in exploratory phase than in planning phase (respectively 2 and 0,62). The inferential analysis shows that this difference is significant: $T_{(18)} = 5,31$; $p=0,0001$.

Table 1: Emotional manifestations according to the solving phases (average and standard deviation)

	Exploratory phase	Planning phase
Number of FE[1] per minute	2,00 *(1,00)*	0,62 *(1,18)*
Average level of SC[2]	1,24 *(0,19)*	1,38 *(0,24)*
Maximum SC amplitude (pic-to-pic)	0,35 *(0,16)*	0,15 *(0,76)*

The average level of skin conductance (table 1) is lower during the exploratory phase than during the planning phase (respectively 1,24 µMho and 1,38 µMho). The inferential analysis shows that this difference is significant: $T_{(18)} = 5,6$; $p=0,0001$.

In return, the maximum skin conductance amplitude (table 1) is higher during the exploratory phase than during the planning phase (respectively 0,35 µMho and 0,15 µMho). The inferential analysis shows that this difference is significant: $T_{(18)} = 6,26$; $p=0,0001$.

3.2 Exploratory phase: Impasse situations versus remaining exploratory phase

The average number of expression facial per minute (table 2) is more frequent during impasses than during the remaining phase (respectively 2,34 and

[1] FE is used for "facial expressions"
[2] SC is used for "skin conductance"

1,43). The inferential analysis shows that this difference is significant: $T_{(17)} = 3,68$; p=0,0009.

Table 2: Emotional manifestations within the exploratory phase (average and standard deviation)

	Impasses	Remaining exploratory phase
Number of FE[1] per minute	2,34 *(1,24)*	1,43 *(0,90)*
Average level of SC[2]	1,21 *(0,18)*	1,27 *(0,19)*
Maximum SC amplitude (pic-to-pic)	0,20 *(0,14)*	0,15 *(0,07*

The average level of skin conductance (table 2) is lower in the impasses than in the remaining phase (respectively 1,21 µMho and 1,27 µMho). The inferential analysis shows that this difference is significant: $T_{(17)} = 3,61$; p=0,0011.

On the other hand, the maximum skin conductance amplitude (table 2) is higher in the impasses than in the remaining phase (0,2 µMho and 0,15 µMho). The inferential analysis shows that this difference is significant: $T_{(17)} = 2,56$; p=0,01.

As predicted, facial expressions of emotion are more frequent in the exploratory phase than in the final phase, and within the exploratory phase they are more frequent in the impasses. Concerning the physiological manifestations of emotion, two apparently paradoxical results were observed. On the one hand, the maximum activity amplitude (pic-to-pic) is higher in the exploratory phase, and within this phase it is higher in the impasses. On the other hand, the average level of skin conductance is lower in the exploratory phase than in the final path, and is lower in the impasses than in the remaining exploratory phase. As a matter of fact, those results are consistent with the hypothesized meanings of the two physiological measures. The higher average level of skin conductance observed in the final phase suggests that it reflects resources mobilisation related to goal achievement proximity, whereas the higher maximum activity amplitude (pic-to-pic) observed in the exploratory phase and in the impasse situations suggests that it reflects the response to emotionally relevant events.

4 Discussion

Those results suggest that the problem-solving situation is emotionally relevant to the solver. Indeed, both facial expressions and the maximum activity amplitude (pic-to-pic) are higher in the exploratory phase, which is actually a problem-solving situation. The subject gets involved in the construction of the relevant interpretation, and must cope with actions sequences that are not relevant in sub-goal achievement. These two criteria are lower in the final phase that is closer to an execution-task. So, it can be argued that it is the problem-solving situation itself that triggers emotional manifestations. In addition, those results suggest that emotion can play a regulatory role in decision-making mechanism. Indeed, both facial expressions and the maximum activity amplitude do not randomly appear in the course of the problem-solving, but in some cognitive significant events like the impasse situations. These two criteria are lower in the remaining exploratory phase and in the final phase, in which actions sequences carried out by the subject are efficient in goal achievement.

On the other hand, the average level of skin conductance is lower in the exploratory phase and, within exploratory phase, is lower when subjects are in impasse situations. This result is consistent with those reported by Pecchinenda and Smith (1993): the difficulty encountered in the exploratory phase temporally leads the subject in a disengagement of the task.

Our results show that the emotional measures used in this research do not reflect the same theoretical constructs. Facial expressions are related to the communication with others, whereas physiological activity reflects the physical preparation for an adequate coping with the situation (Frijda, 1986; Lazarus, 1986; Smith, 1989).

This preliminary research is a first step to integrate the emotional components into the decision-making mechanism described in the Constraints Model (Richard et al., 1993). Our findings show that these components do not randomly appear during the problem-solving, and that their manifestations are linked to specific events of the problem-solving. It shows the necessity to study the emotional manifestations within a cognitive model that allows to describe and to simulate the solver behaviour. In such a model, successes and failures can be identified, and then matched to the emotional manifestations.

5 Conclusion

In conclusion, this research points out the interest to study in a problem-solving situation the emotion as a crucial component in the decision-making mechanism. In this way, we have used a theoretical model that allowed us to identify some cognitive significant events in problem-solving as the successes and failures.

The attempts to experimentally study the emotional manifestations during a problem-solving situation make valuable contributions to our understanding of the general question of the links between emotion and cognition.

Acknowledgements

We express our sincere appreciation to Jean-François Lambert at the University of Paris 8 (France) for his cooperation and our fruitful discussions in this research.

References

J.R. Anderson and C. Lebière. *The atomic components of thought*. Mahwah N.J.: Lawrence Erlbaum, 1998.

R.M. Bauer. Physiologic measures of emotion. *Journal of Clinical Neurophysiology*, 15(5): 388-396, 1998.

R.V. Belavkin. The role of emotion in problem solving. In *Proceedings of the AISB'01 Symposium on Emotion, Cognition and Affective Computing*, pp. 49-57. Heslington, York, England, 2001.

R.V. Belavkin, F.E Ritter and D.G. Elliman. Towards including simple emotions in a cognitive architecture in order to better fit children's behaviour. In *Proceedings of the 1999 Conference of the Cognitive Science Society*, pp. 784. Mahwah, N.J.: Lawrence Erlbaum, 1999.

M.E. Dawson, A.M. Schell and D.L. Filion. The electrodermal system. In J.T. Cacioppo & L.G. Tassinary (Eds.), *Principles of psychophysiology: Physical, social and inferential elements*, pp. 295-324. Cambridge, England: Cambridge University Press, 1990.

P.C. Ellsworth. Some implications of cognitive appraisal theories of emotion. In K.T. Strongman (Ed.), *International Review of Studies on Emotion*, pp. 143-161. New York: Wiley, 1991.

N.H. Frijda. *The emotions*. Cambridge: Cambridge University Press, 1986.

S.C. Gadanho and J. Hallam. Robot learning driven by emotions. *International Society for Adaptative Behavior*, 9(1): 42-64, 2002.

J. Gratch and S. Marsella. A domain-independent framework for modelling emotion. *Cognitive Systems Research*, 269-306, 2004.

S. Kaiser and T. Wehrle. Facial expressions as indicators of appraisal processes. In K.R. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research*, pp. 285-300. Oxford University Press, 2001.

K. Kotovsky and D. Fallside. Representation and transfer in problem solving. In K. Kotovsky (Ed.), *Complex information processing (What has Simon brought?). 21st symposium of the Carnegie Mellon Institute*, pp. 69-108. Hillsdale, N.J.: Erlbaum, 1989.

R.S. Lazarus. Emotions and adaptation: Conceptual and empirical relations. In W.J. Arnold (Ed.), *Nebraska Symposium on Motivation*, 16: 175-226. Lincoln, NE: University of Nebraska Press, 1986.

R.S. Lazarus. *Emotion and adaptation*. New York: Plenum Press, 1991.

K. Oatley and P.N. Johnson-Laird. Towards a cognitive theory of emotions. *Cognition and Emotion*, 3: 125-137, 1987.

A. Pecchinenda and C.A. Smith. The motivational significance of skin conductance activity during a difficult problem-solving task. *Cognition and Emotion*, 10: 481-503, 1996.

J.F. Richard. Comportements, buts et représentations. *Psychologie Française*, 44(1) : 75-90, 1999.

J.F. Richard, S. Poitrenaud and C.A. Tijus. Problem-solving restructuration: Elimination of implicit constraints. *Cognitive Science*, 17 : 497-529, 1993.

J.F. Richard and M. Zamani. A problem-solving model as a tool for analysing adaptative behavior. In R.J. Sternberg, J. Lautrey & T. Lubart (Eds.), *Models of intelligence: An international perspective*. Washington, DC: American Psychological Association, 2003.

I.J. Roseman. Appraisal determinants of discrete emotions. *Cognition and Emotion*, 5: 161-200, 1991.

D. Sander and O. Koenig. No inferiority complex in the study of emotion complexity: A cognitive neuroscience computational architecture of emotion. *Cognitive Science Quarterly*, 2: 249-272, 2002.

K.R. Scherer. On the nature and function of emotion: A component process approach. In K.R. Scherer & P. Ekman (Eds.), *Approaches to emotion*, pp. 293-317. Hillsdale, NJ: Lawrence Erlbaum Associates Inc., 1984.

K.R. Scherer, A. Schorr and T. Johnstone. *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press, 2001.

H.A. Simon. Motivational and emotional controls of cognition. *Psychological Review*, 74(1): 29-39, 1967.

C.A. Smith. Dimensions of appraisal and physiological response in emotion. *Journal of Personality and Social Psychology*, 56: 339-353, 1989.

C.A. Smith and R.S. Lazarus. Emotion and adaptation. In L. A. Pervin (Ed.), *Handbook of personality theory and research*, pp. 609-637. New York: Guilford, 1990.

Emotions as Evaluations

Peter Goldie and Sabine A. Döring
Department of Philosophy
King's College London
Strand
London WC2R 2LS
United Kingdom
peter.goldie@kcl.ac.uk, sabine.doering@kcl.ac.uk

Abstract

We should distinguish between two kinds of psychological states which are often not properly distinguished: desires and evaluations. Emotions necessarily involve evaluations but don't necessarily involve any immediate or direct connection with desires. This distinction throws light on the ways in which human emotional experiences can play diverse and complex explanatory roles. If an artificial agent is to be constructed to be emotionally anything like a human biological agent, then there must be no immediate or direct connection between emotion and motivation understood as desire. An implication of our approach is that an artificial or non-human biological agent with the relevant desires is not yet an agent with emotion. Moreover, with the above distinction in place it becomes less easy to see how to construct an agent capable of emotion—that is, an agent capable of emotional evaluation without any immediate or direct connection to desire.

1 Introduction

We should distinguish between two kinds of psychological states which are often not properly distinguished. First, there are those motivating states that are standardly taken to be necessary to explain action. These states we will call desires. Secondly, there are those states that are necessarily involved in emotion, but which have no direct role in explaining action and which have no immediate or direct connection to desire or to motivation in general. These states we will call evaluations. Emotions necessarily involve evaluations but don't necessarily involve any immediate or direct connection with desires.

This distinction throws light on the ways in which human emotional experiences can play diverse and complex explanatory roles. If an artificial agent is to be constructed to be emotionally anything like a human biological agent (using the terms suggested by the organisers of the symposium), then there must be no immediate or direct connection between emotion and motivation understood as desire. One might think that the mark of a successful construction of emotion in an artificial or non-human biological agent would be that the agent has the relevant desires. But an implication of our approach is that an artificial or non-human biological agent with the relevant desires is not yet an agent with emotion. Moreover, with the distinction between emotional evaluation and desire in place, it becomes (as it clearly should become) less easy to see how to construct an agent capable of emotion—that is, an agent capable of emotional evaluation without any immediate or direct connection to desire.

2 Human action and motivation

Let's begin with human action. Actions, as such, aren't just bodily movements, for bodily movement is neither necessary nor sufficient for action: it's not necessary because I can do something without my body moving (such as signalling to someone by sitting dead still); and it's not sufficient because my body might move without my doing any action (such as my trembling with cold, or my knee moving when it's hit by a doctor's mallet).

Although there are many instances which are clearly and uncontroversially either actions or just bodily movements, there are controversial instances where it is not clear where action ends, and where other kinds of 'behaviour' begin: for example, many

of our expressions of emotion—facial expressions (grimacing with anger), and expressive behaviour (such as kicking the chair in anger). But to make progress, let's begin with human actions which are most uncontroversially actions, where an action is understood, roughly, as something done by someone intentionally[1]. This would include actions not done out of emotion (such as going to meet someone because you promised), and actions done out of emotion (such as, in anger, hitting the person you are angry with).

What kind of motivation is necessary for action, so to speak lying behind the intention? One might say it is desire, so that whatever I do or try to do intentionally, I do because I desire or want the thing that I think my action will bring about[2]. But if this is to be accepted, then the notion of desire will have to be very protean. One reason for this is that it is often possible for us to do things that, just in terms of the phenomenology of desire (that is, desires which have a certain 'feel'), we have no desire to do. Even in respect of actions out of emotion, motivation often lacks this phenomenology—a phenomenology which is 'traditionally distinguished from "cold" cognition' in the words of the symposium organisers. For example, the Mafia advise us that, if one is angry with someone and wants revenge, then this revenge is a dish best tasted cold. In respect of some actions, the person acting might not just deny that he has a desire with this phenomenology; he might even deny that he has any desire at all to do what he does. For example, one evening you go to meet someone whom you earlier promised to meet, and you might sincerely say that you went because you promised to go, and not because you wanted to; all you really wanted to do, you say, was to stay in and watch the football on TV. So, if we are to find room for examples like these, we must have a more protean notion of motivation than desire in the phenomenological 'feel' sense, and more protean than desire of the kind which the person acting is conscious of[3].

[1] To be a bit less rough, an action is something done by someone which is intentional under a description. So someone might intentionally be hammering in a nail but unintentionally waking the neighbours. This idea of action and intention is discussed in the seminal work of Anscombe (1957), and in Davidson (1963).

[2] As Anscombe puts it (1957: 67), with animals in mind, '[T]he primitive sign of wanting is trying to get'.

[3] The role of desire in action explanation is discussed in Nagel (1979), Smith (1987; 1994), and Schueler (1995).

2 Pro-attitudes and means-end reasoning

At this point, the notion of a 'pro-attitude' is often introduced to stand duty for the notion of desire, with the thought that pro-attitudes should include states that are, in respects to be discussed below, like desires, but are not necessarily desire-like in their phenomenology or in their being states of which one is necessarily conscious. The philosopher Donald Davidson, for example, has characterised pro-attitudes as including 'desires, wantings, urges, promptings, and a great variety of moral views, aesthetic principles, economic prejudices, social conventions, and public and private goals and values …' (1963:4).[4]

So far, then, we have the theory that a pro-attitude is a supposedly necessary motivation for action. But it is not a sufficient motivation for action. First, one might have a pro-attitude towards an action, but this pro-attitude is 'outweighed' by some other pro-attitude towards some other action, where both actions cannot be performed (you could have a pro-attitude this evening towards watching the football on TV and a stronger pro-attitude towards keeping your promises, in which case, presumably, you will not watch the football on TV). Secondly, and more interestingly for our purposes, a pro-attitude is not sufficient for action because there also needs to be some other kind of psychological state—paradigmatically some kind of belief or perception—that represents the way the world is and that represents your action as the 'means' to bringing about your 'end'—to bringing about whatever it is you have a pro-attitude towards. This is what lies behind the familiar idea of means-end reasoning. For example, your having a pro-attitude towards drinking some beer isn't sufficient to explain your drinking from this glass; you also need to see that what's in this glass is beer, and to believe that your moving the glass in this way is the suitable means to your end of drinking some beer.

So we now have developed our theory to this: actions—all actions—can be explained by, and are brought about by, two distinct kinds of psychological state, distinguished by their characteristically different roles in explaining and bringing about action: pro-attitudes, which roughly represent the way the world would be if the pro-attitude is satisfied; and beliefs, which roughly represent the way the world is, and the means of changing the world so that the pro-attitude is satisfied.

[4] Davidson himself would seem to think that all pro-attitudes 'can be interpreted as attitudes of an agent directed toward actions of a certain kind'; as will emerge, it is this connection that we dispute.

This may seem like a complex bit of theorising, but the idea—at heart just means-end reasoning—really isn't too far from our common sense way of explaining action. Of course in explaining everyday action, we sometimes just refer to the pro-attitude, and sometimes just to the belief, but this will typically be for pragmatic reasons. If you ask me why I crossed the road, I might reply 'To get a sandwich', which reply presupposes that you believe, as I do, that there is a sandwich shop across the road, so all I need to refer to is my pro-attitude. But if I reply 'Because I think there is a sandwich shop across the road', this presupposes that you know that I want a sandwich, but that the presence of the sandwich shop is for some reason in doubt, so I need to refer to this belief to explain my action, but not to the pro-attitude.

Let's assume for the purposes of this paper that this theory is correct, and that all actions are brought about, and are explicable, by a pro-attitude and a belief[5]. It would follow that action out of emotion must be explicable in this way. Thus, according to this picture, my action of hitting you out of anger must be explicable by reference to a pro-attitude, such as my pro-attitude towards getting my revenge for the wrong you have done me, and to a belief, such as my belief that hitting you is a suitable and available means of getting my revenge.

3 Pro-attitudes: desires and evaluations

The problem now is that the notion of pro-attitude that we have before us would seem to be too protean, for many pro-attitudes are such as never to lead to action of any kind, even with the relevant belief in place. Why not? Because (we are assuming) all action needs a desire in place, and not all pro-attitudes are desires. For example, I might have a negative attitude (a 'con-attitude') towards the presence of chewing gum on the pavements of London, and yet do nothing about it, because I have no desire to do anything about it. Or I might have a negative attitude towards you, but never do anything to harm you in any way, again because I have no desire to.

What we need, it seems, is the kind of distinction that the notion of pro-attitude has blurred: a distinction between, on the one hand, those kinds of pro-attitude which are desires, and which do imply a necessary and direct motivation towards action (but which do not necessarily have a phenomenology or involve any conscious awareness); and, on the other hand, those kinds of pro-attitudes which we will call evaluations, and which imply no immediate or direct motivation towards action.

Armed with this distinction, we can readily see that there can be evaluation without desire. Putting the point roughly, there are all sorts of things that we humans care for or value, sometimes for a whole lifetime, but which we never do anything to bring about or to promote. One would have to be in the grip of a theory to insist that the fact that you do nothing to bring something about or to promote it shows that you don't care about that thing or value it in any way.

Another way to think of the point might be this. There might be a necessary connection from (1) acting to bring something about, to (2) desiring to bring that thing about, to (3) valuing (at least in some etiolated sense) the thing that you want to bring about. But there is no necessary connection from (3) valuing something or having a pro-attitude towards its coming about, to (2) desiring to bring that thing about, and thus to (1) acting to bring that thing about. There is a fundamental distinction between 'ought to be', as an expression of evaluation, and 'ought to do', as an expression of desire.

4 Emotions: desires and evaluations

How does all this bear on emotion and motivation? Well, we can now begin to see good reasons for thinking that emotions necessarily involve evaluations, but do not necessarily involve desires: and my anger at the chewing gum on the pavements is an example where evaluation and desire come apart. Then what this thought does—and this is the advantage as we see it of our approach—is to shift our focus from the very intractable and oblique question of what the connection is between emotion and 'motivation', understood as desire. Part of the problem here, which makes the question so intractable and oblique, is that emotion and desire are doubly dissociable: one can have emotion without desire; and one can have desire without emotion. Instead, we can move to the more tractable questions of what the connection is between, on the one hand, emotion and evaluation and, on the other hand, between emotional evaluation and desire.

Why should we think that there is a necessary connection between emotion and evaluation? This is too big a question for us to address in this short pa-

[5] The theory is in fact highly contentious; see, for example Döring and Peacocke (2001) and Döring (2003), who bring it into question. Explaining action might be one thing, and bringing action about or causing action something else. But for the purposes of this paper, it is best to leave these issues to one side.

per, but the underlying idea is really very simple[6]. Emotions—all emotions—are concerned with what we care about or what we value. Let's consider, for example, some historical fact, such as the fact that the treatment of slaves in eighteenth century Europe was unjust, or some future possibility, such as the possibility that there will be no whales remaining on the planet in two hundred years' time. Now I might just believe these facts without caring about them: I believe that slaves were treated badly in those days, and I believe that it's possible that fairly soon there will be no whales on the planet. But what if I feel angry about the unjust treatment of slaves, and what if I fear the extinction of the whale? Now I care about these facts, rather than just believing them to be true.

But will I do anything about these things? Will I desire to do anything about these things? Our answers to these questions is 'Maybe, maybe not'. Of course, in respect of many emotions, there is evaluation and desire and action, as would be the case with my immediate fear of the wild bull's charge: when I run away from it out of fear, I both evaluate the bull as fearful, and I desire to get away from it as the immediate object of my fear. But it is a mistake to think of all human emotion as being just like this kind of example: not all emotions are 'affect program responses'—short-term, visceral responses such as fear of the wild bull's charge.[7]

Indeed, a little thought will reveal that many of our emotions are evaluative concerns of the kind which do not involve any immediate or direct connection with action or with desire. We have emotions about the past and about the future, where we have no related desires—recall the earlier examples of anger at the slavery, fear for the whales. We feel emotions about things we remember doing, such as shame at that needlessly cruel thing we did last year, but we don't desire to do anything about it—it's too late for that.[8] We get emotionally involved with fictional characters, caring whether or not the heroes of the Western we are watching will triumph over the corrupt Marshall; and yet we do not try to intervene on their behalf.[9] We value artworks highly, are moved by them emotionally, thinking them as objects of great beauty perhaps, but we don't necessarily have any desires for them: to think that we do would be to turn all aesthetic appreciation of artworks into a desire to collect them or to own them.[10] We feel awe at the stars in the heavens or at the storm waves crashing against the sea wall, but we have no relevant desires; yet this feeling, known by Kant and Burke in the eighteenth century as a feeling of the sublime, is surely a kind of emotion (Kant: 1790/1987; Burke 1757). We put ourselves in someone else's shoes and imagine how they would feel (such as imagining being attacked in a dark alleyway); empathetically we can come to feel real emotions, but we do not desire to act on those emotions—we feel real fear at what we imagine, but we don't desire to run away from what we imagine.[11] All emotions are evaluations, but by no means do all emotions involve any immediate or direct connection with desire. So we claim.

5 How one might try to refute our claim

It is always open to our opponent to insist that emotions always involve desire. We will first consider an example, and then we will suggest some ways that our opponent might resist our claim.

Let's go back to our example of my anger at the presence of chewing gum on pavements of London. This is a real emotion. But I do nothing about it. Our claim is that it is possible that the reason why I do nothing about it is that I have no desire to do anything about it: I might have the thought 'Something ought to be done about this', or 'The pavements ought not to look like this and feel so sticky underfoot', but I needn't also have the thought 'I ought to do something about this'. Our opponent denies this: she claims that emotion implies evaluation (as we do), and that emotional evaluation implies an immediate and direct necessary link to desire (as we deny). Then our opponent goes on to list a wide range of reasons why we don't act in this case, which maintain an immediate and direct link between emotional evaluation and desire. Here are some:

- She might accept that I care, and insist that I do desire to do something; but the reason I don't act is that my desire to do something is outweighed by a stronger contrary desire (perhaps a prudential

[6] For detailed discussion, see Nussbaum (2003) and Solomon (2003).

[7] For discussion of affect program responses, see, for example Griffiths (1997), and Ekman (1992).

[8] For discussion of our emotional responses to remembered events, see Goldie (2003).

[9] This is an example of the so-called paradox of fiction, much discussed by philosophers. See for example Currie (1990) and Walton (1990).

[10] As John McDowell says, in criticising J. L. Mackie's notion of objective prescriptivity, such a notion 'suggests a specific response involving value's appeal to the *will*, and this is at best questionably appropriate for ethical value in general, and surely inappropriate for aesthetic value' (1998: 117).

[11] For discussion of the role of emotion in simulation, see Goldie (2002).

one to conserve my energy and resources).

- She might insist that there is always some kind of action, such as verbal action ('Something ought to be done'), or expressive action, including expressive verbal behaviour (such as a sigh of despair) and expressive gestures. So the fact that there is always action shows that, after all, desire of some kind or other is always present.
- She might insist that the kind of desire that is present could be a dispositional desire for later action, rather than a desire for immediate action; in this example, I might desire to vote for the politician who pledges to remove chewing gum from the pavements.
- She might say that the connection between evaluation and desire is a normative one, so that a failure to form a desire is a failure of rationality of some kind; in this example, I might fail to desire to do anything, in spite of my emotion, because of sloth, or what the ancient scholastics called accidie.

How should we respond to this battery of possibilities, each of which has a venerable philosophical history? First, we should accept that each of these possibilities is at least possible in individual cases. Secondly, we should accept that we don't have a knock-down argument to show that desire of some kind or other needn't always be present. But thirdly, we should, like in judo, use our opponent's claims against her: by putting forward this battery of possibilities, she has shown just how deferred and indirect the connection between emotional evaluation and emotional desire can be in us humans. One can be beguiled by simple examples, such as the wild bull, and brought to think that evaluation (of the bull as fearful) and desire (to get away from what is fearful) are always immediately and directly linked, or even that evaluation and desire are really just two sides of the same coin. Fundamentally, these are the thoughts that we want to undermine. Emotional evaluation and emotional desire must be seen as distinct.

6 Implications

What are the implications of this for the topic of the symposium and for those who are concerned with the construction of 'agents' with emotions?

Imagine that one built an artificial agent that was capable of recognising warm spots on the floor; on recognising a warm spot, it would move towards it, overcoming obstacles on the way, and then settle there until the spot cooled down. Then it would recognise other warm spots in the vicinity, and repeat its behaviour. Could we say that one had built an agent which was capable of emotion? Surely not. Surely all that one has is an agent whose movements might be characterised as explicable, and brought about by, means-end reasoning—by a combination of desire and belief or perception: it desires to be on a warm spot; it recognises that there is a warm spot in some direction or other, and it believes that the best means of getting to the warm spot is to move in such-and-such a direction and in such-and-such a manner. (We should really have scare quotes around 'desire' and 'belief', but leave that to one side.) We have no reason to assume that there is, in the agent, any kind of emotional evaluation of warm spots, and, moreover, the principle of Occam's Razor (that one should not multiply entities beyond necessity) suggests that we would be wrong to do so.

So how could one build an artificial agent with emotion? The connection between emotional evaluation and desire should not be too direct and immediate. All emotions are not like hunger or thirst or itches or fear of the wild bull. Rather, the agent should be constructed to be capable of manifesting a wide range of connections between emotional evaluation and emotional desire in individual cases, leaving a significant place for the battery of possibilities put forward by our opponent earlier. In short, it seems a relatively easy task to build an artificial agent with 'desire', but it seems (as it should seem) like no easy matter to build an artificial agent with 'emotional evaluation', where this is clearly distinct from desire in the way that we have been insisting on: in effect one has to build an agent that is capable of the thought 'ought to be' as distinct from the thought 'ought to do'.

Now let's turn to 'biological agents'. Our remarks about evaluation and desire reveal how great is the gap between non-human biological agents (dogs, octopi, rats) and human biological agents. Maybe, in such creatures, one can make do with a more immediate and direct connection between emotional evaluation and desire, as if all emotions are like hunger, thirst and so on; in other words these creatures more readily fit the paradigm of the example of fear of the wild bull, where the link between evaluation and desire is of that kind. But humans are not like that: with us, evaluation and desire need have no such link.

Finally, we should mention the considerable ethical implications of our view. If humans really are that different emotionally from other biological and artificial agents, and the connection between

emotional evaluation and desire is that much more tenuous and oblique, then methods of measuring emotion as an indicator of motivation, understood as desire, that might be appropriate for other agents may well not be appropriate for humans. I could be very angry with you, but have no desire to do anything about it.

Acknowledgements

This work is supported by the EU-funded Network of Excellence HUMAINE, FP6-IST contract 507422. The views expressed here are those of the authors and not necessarily those of the consortium.

References

Elisabeth Anscombe. *Intention*. Oxford: Blackwell, 1957.

Edmund Burke. *A Philosophical Enquiry into the Origin of our Ideas of the Sublime and Beautiful*, ed. J. T. Boulton, London, 1757.

Greg Currie. *The Nature of Fiction*. Cambridge: Cambridge University Press, 1990.

Donald Davidson. Actions, Reasons, and Causes. *Journal of Philosophy* 60, 1963. Reprinted in his *Essays on Actions and Events*. Oxford: Oxford University Press, 1980; page numbers refer to this.

Sabine Döring. Explaining Action by Emotion. *The Philosophical Quarterly* 211, 214-30, 2003.

Sabine Döring und Christopher Peacocke. Handlungen, Gründe und Emotionen. In *Die Moralität der Gefühle*, eds Sabine Döring and Verena Mayer, Special Issue of *Deutsche Zeitschrift für Philosophie*, Berlin: Akademie, 81-103, 2002.

Paul Ekman. An Argument for Basic Emotions. *Cognition and Emotion* 6, 169-200, 1992.

Peter Goldie. Emotion, Personality and Simulation. In *Understanding Emotions: Mind and Morals*, ed. Peter Goldie, Aldershot: Ashgate Publishing, 2002.

Peter Goldie. One's Remembered Past: Narrative Thinking, Emotion, And the External Perspective. *Philosophical Papers* 32, 301-19, 2003.

Paul Griffiths. *What Emotions Really Are: The Problem of Psychological Categories*. Chicago: University of Chicago Press, 1997.

Immanuel Kant. *The Critique of Judgement*. Tr. W. S. Pluhar, Indianapolis: Hackett, 1790/1987.

John McDowell. Aesthetic Value, Objectivity, and the Fabric of the World. Reprinted as Chapter 6 in his *Mind, Value, and Reality*, Harvard: Harvard University Press, 1998.

Thomas Nagel. *The Possibility of Altruism*. Princeton: Princeton University Press, 1979.

Martha Nussbaum. *Upheavals of Thought*. Cambridge: Cambridge University Press, 2003.

George Schueler. *Desire: Its Role in Practical Reason and the Explanation of Action*. Harvard: MIT Press, 1995.

Michael Smith. The Humean Theory of Motivation. *Mind* 96, 36-61, 1987.

Michael Smith. *The Moral Problem*. Oxford: Blackwell, 1994.

Robert Solomon. *Not Passion's Slave*. Oxford: Oxford University Press, 2003.

Kendall Walton. *Mimesis as Make-Believe*. Harvard: Harvard University Press, 1990.

A consideration of decision-making, motivation and emotions within Dual Process theory: supporting evidence from Somatic-Marker theory and simulations of the Iowa Gambling task.

Kiran Kalidindi[⋆]
[⋆]Centre for Cognitive Neuroscience and Cognitive Systems
Computing Lab, Univ. of Kent
Canterbury, Kent, UK. CT2 7TR
kk49@kent.ac.uk

Howard Bowman[⋆†]
[†] H.Bowman@kent.ac.uk

Brad Wyble[⋆‡]
[‡] B.Wyble@kent.ac.uk

Abstract

For many years there have been lay people, philosophers and scientists who have made the distinction between affect and 'cold' cognition. This paper examines the potential value of this dichotomy in relation to understanding ventromedial lesion (VMF) patients behaviour in general and on the Iowa Gambling Task (IGT). We use a combination of dual-process and somatic-marker theories and inferences based on simulations of normal controls and VMF patients on the IGT.

1 Introduction

Two closely related dichotomies. In recent years some researchers in the field of thinking and reasoning have been proposing dual-process accounts for the 'non-rational' results of human performance on normative logical tasks. The main evidence comes from deductive reasoning paradigms, particularly the Wason selection task, both abstract and deontic versions, the belief bias found in syllogistic reasoning (Sloman, 1996) and neuropsychological data (Goel and Dolan, 2003). These dual-processes are said to emanate from two quite separate cognitive systems that have distinct evolutionary histories. "System 1 is old in evolutionary terms and shared with other animals: it comprises a set of autonomous subsystems that include both innate input modules and domain-specific knowledge acquired by a domain-general learning mechanism. System 2 is evolutionarily recent and distinctively human: it permits abstract reasoning and hypothetical thinking, but is constrained by working memory capacity and correlated with measures of general intelligence." (Evans, 2003). This dual-process split is similar to another often mentioned dichotomy of affect and 'cold' cognition, where traditionally, motivation and emotions are considered to be part of affect.

Such dichotomies have long been a staple for philosophers and psychologists alike, stemming from Plato and Aristotle to James (1890) and to current work in journals like Cognition and Emotion. By placing affect and 'cold' cognition within the context a seemingly similar theory, like dual-process theory, it becomes possible to further constrain our models and move towards an elucidation of what emotions are, how emotions might help in reasoning by seeing when they occur and what causal effect they may have. These are particularly important issues since currently we have no clear accepted definition of what phenomena emotions include (Evans, 2002; Sloman, 2004).

Bringing the two dichotomies together. These two dichotomies seemingly represent the same functional split, but are not currently considered together within the literature. A key question concerning the systems in these dichotomies is how each system interacts with one another. Do they have different memory systems? What might be the method of internal communication between them? Can they interact individually with the external world? Does one system become more dominant than the other in certain types of situation and/or context? A useful starting point in such research can be to look at a theory that seems to represent the affect/'cold' cognition dichotomy and see what happens when the systems become somewhat disconnected, i.e. when they work with limited support from the other type of processing. We will argue that such a separation arises in VMF patients.

2 Somatic-Marker Theory and the Iowa Gambling Task

One theory we can examine is the idea of somatic-markers. It proposes that body states act as a valence that can be associated with potential choices based on prior outcomes, and thus aid decision-making, both by limiting the search space and allowing an affective evaluation of choices. The main supporting evidence for this theory arose from clinical interviews of subjects with ventromedial prefrontal cortex (VMF) lesions and their performance on the Iowa Gambling Task (IGT) (Bechara et al., 1999), compared to normal controls and those with lesions in other brain areas. The gambling task consists of four decks that subjects can pick from; two decks, A and B, which yield high wins but higher losses (Disadvantageous) and the other two, C and D, that yield low wins with lower losses (Advantageous). Normal subjects start by picking from the disadvantageous decks but learn to pick from the advantageous decks, unlike the VMF patients who, as in their real social and personal lives, continue to pick non-advantageously.

Going back to the potential dichotomy of affect and 'cold' cognition, it has been suggested by Damasio (1994) that patients with VMF lesions are no longer able to automatically produce somatic-markers when making social decisions. Unlike laboratory experiments, such decisions often have large numbers of choices and unlimited consequences. Some somatic-markers can be considered synonymously with 'feelings' as the conscious component of emotions. VMF patients possess most of their 'cold' cognitive faculties, as shown by normal verbal test, tower of hanoi and IQ scores, but seem to have blunted emotions when describing situations, even when they are intimately involved. Parts of the affect system are still intact in VMF patients, as they can be classically conditioned and they get SCR responses when they are rewarded and punished in the IGT (Bechara et al., 1999). This could suggest that the VMF region is where social event knowledge is held (Wood et al., 2003), and is an important link between the 'cold' cognitive system (explicit memory) and the affect (emotional memory) system. Alternatively, is it just an important part of a single overall system? There are often many interpretations to data concerning the brain, mind and behaviour. However, one way forward is to propose hypotheses and then see how they stand-up to current experimental results and to new direct hypotheses tests, i.e. following the Popperian view of scientific discovery (Popper, 1959).

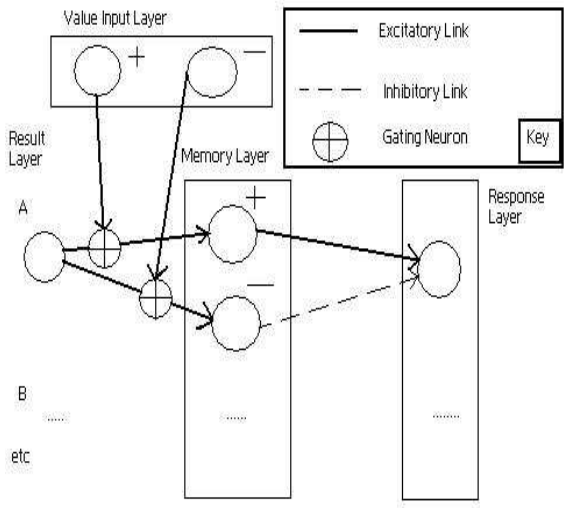


Figure 1: **Diagram of Neural Network used for the Iowa Gambling Task. It shows the repetition of the basic architecture for each choice in the problem space.**

3 Neural Network Model of the IGT

A dual system account for VMF behaviour arose from the current literature and neural network simulations examining the different behaviours of normal controls and VMF patients on the Iowa Gambling Task (IGT), at a level of abstraction above and, in this case, inclusive of both the 'cold' cognitive and affect systems (Kalidindi et al., 2005). An explanation that accounts for the difference in behaviour between normal controls and VMF patients on the IGT is 'myopia' for future consequences, in that they are driven by immediate reward and are less interested in uncertain future loss or gain (Bechara et al., 2000b). This simulation investigates the implications of this 'myopia'. The current literature lacks a model that both accurately reproduces these experimental results, and is abstracted from specific anatomical details, which underlie other models (Wagar and Thagard, 2004). A diagram of the neural network can be found in Figure 1. It shows the Memory Layer where the affinity for each choice in the choice space is represented by a positive and negative pair of units.

The simulation begins by using a random number generator to pick one of the four decks, this choice/result is represented by an activation value of 1 in the relevant Result Layer unit. Then a card is picked from the chosen deck and the result is fed into the Value Input Layer, where the positive neu-

ron is activated by money won and the negative unit by money lost. The activation is simply the amount divided by $2000 (the amount of money the player starts with) or the largest card result seen so far, if that is larger. Through gating neurons (McClelland and Rumelhart, 1986) this activation is passed onto the relevant Memory Layer pair (See Figure 1) by multiplying the activation in the Result Layer units with that from the Value Input Layer; as only one Result Layer unit is active, only the units associated with that choice will recieve new activation. Next, the Response Layer units will recieve excitatory input from the linked positive Memory Layer unit and inhibitory input from the linked negative unit. Then a normally distributed random activation is added to each Response Layer unit to encourage exploration. This is followed by a winner-takes-all competition, where the winning unit in the Response Layer becomes the deck choice for the next trial, by activating the relevant Result Layer unit and returning its' own actiation to zero. This process continues through the trials of the task, with a build-up of knowledge about each deck being held in the Memory Layer units. (Note, all the units, except those in the Memory Layer and the active unit in the Result Layer (representing the next choice), are set to zero after each trial.)

To explore the difference between normal controls and VMF patients, a time-averaging equation (See Figure 2) was used to describe the decay of information, and how new and old information influences Memory Layer units. This relationship between old and new information can be altered by changing the parameter τ, which is a real number between 0-1. Thus, if τ is close to 1 then previous activation (i.e. information) is almost completely preserved while current activation (i.e. information) has little impact on the representation of the valence of a choice. It is felt that VMF patients might be less able to integrate past information into current decision-making than normals. We suggest that this might be the cause of VMF patients' 'myopia' for future consequences. An exploration of the state space of τ was performed to confirm which values of τ best matched the normal control and VMF patient behaviour on the IGT.

The results showed that a constant value for τ, across cycles, equal to 0.52, gave the closest deck choice profile to that of the VMF patients (a 2% difference between the simulated and the human data across the 5 data points, that occur every 20 selected cards, (Bechara et al., 2000b)), whereas increasing τ over trials was required to reproduce the normal controls deck choice profile. Various increasing functions over cycles gave almost exact replicas of the

$$act(t) = \tau \cdot act(t-1) + (1-\tau) \cdot y_\beta$$

Figure 2: **Time-Averaging Equation - act(t) is the activation in the unit at time t, τ is the time-averaging parameter and y_β is the gated input from the Value Input Layer.**

normal human subjects profiles (a 6% difference between simulated and real data). Therefore, it seemed that an increase over trials was the key point. Overall, this could suggest that VMF patients always treat any new trial information as they would treat the first trial in the task, suggestive of their inability to progress from disadvantageous decks to advantageous decks in the IGT. However, in normal subjects, as experience over trials increases, less value is put on current-input (information/activation) and more value is placed on holding onto past experience. The next question for investigation could be what are the underlying causes of this?

4 Conclusions

Implications of the model. Our modelling work suggests that (normal) humans have an effective mechanism by which the weighting applied to past vs present information changes during an "exploration" task, such as the IGT. Furthermore, it may in fact be the case that such strategic adjustments are central to optimum decision-making, especially decision-making in non-initial phases of such "exploration" tasks. The results further suggest that VMF patients have lost this capacity to strategically adjust this past vs present weighting as an "exploration task" proceeds. Our finding here is largely consistent with the 'myopia' for future consequences theory of VMF patient behaviour (Bechara et al., 2000a).

It is well accepted that the ventromedial prefrontal region is implicated with emotion and body-state influences on decision-making. Somatic-markers are one theory of such affective influences on decision-making. If we accept the somatic-marker theory, our modelling results suggest that somatic-markers play a particularly significant role in encoding, maintaining and utilising (in future decisions) past information and further, that such somatically-driven memory and retrieval is particularly important in post initial stages of an "exploration task". Indeed, such mechanisms may become progressively more important through-

out the time course of such a task. Such strategic adjustment of past vs. present weighting during exploration also has relevance for the construction of artificial agents, especially those that seek to take inspiration from somatic-marker theories of human decision making.

General Implications and Proposals. More generally and speculatively, it could be proposed that dual process theory's System 1 and the affect system can be considered as largely synonymous. We propose that the goals, and therefore System 1 alone, are motivated by fairly basic needs, such as quenching thirst, satiety of hunger, obtaining what is envied and sexual encounters. This is in-line with the proposal that System 1 is shared most closely with animals.

If we suggest that a main cause of VMF behaviour is due to a weakening of the connection between System 1 and System 2 and that System 1 normally has greater influence during social and personal situations, then, as System 2 has even less influence over System 1 in VMFs than in normal subjects, we would expect VMF patients to become almost whimsical in social, personal or uncertain situations, and that System 2 responses might even be ignored. This lack of influence by System 2 is shown in the IGT when, normals (70%) and some VMF patients (50%) gain conceptual (conscious) knowledge (Bechara et al., 2000a), by around the eightieth cycle/card, of which decks are advantageous. But unlike normal controls the VMF patients fail to use this knowledge and do not improve their deck choices/strategy. A further cause for VMF behaviour is, during the build-up of knowledge about the IGT, normal subjects develop representations or access to representations of explicit events and their associated affect content through the VMF. This claim is supported by the VMF's known reciprocal connections with the amygdala (important in affect) memory and the hippocampus (important in episodic memory). VMF patients do not gain a build-up of this knowledge and therefore, have a constant τ. Our simulations would also suggest that VMF regions in normals reduce the impact of new information about a situation as the associated information/experience increases. This combination of explicit (System 2) and affect (System 1) memory is used to 'hypothesis' test against the expected outcomes of a choice. Access to this combination of System 1 and System 2 information can be considered equivalent to the effective use of somatic-markers. Therefore, the VMF regions provide a representation link between System 1 and System 2, allowing for an 'affect' assessment of the combination of potential reward and punishment outcomes for an explicit choice, without the need for explicit rules.

References

A. Bechara, A.R. Damasio, H. Damasio, and G.P. Lee. Different contributions of the human amygdala and ventromedial prefrontal cortex to decision-making. *The Journal of Neuroscience*, 19(13):5473–5481, 1999.

A. Bechara, Damasio H., and Damasio A.R. Emotion, decision making and the orbitiofrontal cortex. *Cerebral Cortex*, (10):295–307, 2000a.

A. Bechara, D. Tranel, and H. Damasio. Characterization of the decision-making deficit of patients with ventromedial prefrontal cortex lesions. *Brain*, (123):2189–2202, 2000b.

A.R. Damasio. *Descartes Error: Emotion, Reason and the Human Brain*. New York: Grosset/Putnam, 1994.

D. Evans. The search hypothesis of emotion. *British Journal of Philisophical Science*, (53):497–509, 2002.

J.St.B.T. Evans. In two-minds: dual-process accounts of reasoning. *Trends in Cog Sci*, 7(10):454–459, 2003.

V. Goel and R.J. Dolan. Explaining modulation of reasoning by belief. *Cognition*, (87):B11–B22, 2003.

W James. *The principles of psychology*. New York: Dover, 1890.

K. Kalidindi, Bowman H., and Wyble B. An investigation of the myopia for future consequences theory of vmf patient behaviour on the iowa gambling task; an abstract neural network simulation. In *Proceedings of the Neural Computation and Psychology Workshop 9*, 2005.

J.L. McClelland and D.E. Rumelhart. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations*. MIT Press, 1986.

K.R. Sir. Popper. *The logic of scientific discovery*. London : Hutchinson, 1959.

A. Sloman. What are emotion theories about? Invited talk at AAAI Spring Symposium, March 2004. See - http://www.cs.bham.ac.uk/research/cogaff/sloman-aaai04-emotions.pdf.

S.A. Sloman. The empirical case for two systems of reasoning. *Psychological Bulletin*, 19(1):3–22, 1996.

B. Wagar and P. Thagard. Spiking phineas gage: A neurocomputational theory of cognitive-affective integration in decision making. *Psychological Review*, (111):67–79, 2004.

J.N. Wood, S.G. Romero, M. Makale, and J. Grafman. Category-specific representations of social and nonsocial knowledge in the human prefrontal cortex. *Journal of Cognitive Neuroscience*, 15(67), 2003.

Emotion and motivation in embodied conversational agents

Nicole C. Krämer*
*Department of Psychology
University of Cologne
Bernhard-Feilchenfeld-Str. 11
50969 Köln
nicole.kraemer@uni-koeln.de

Ido A. Iurgel†
†Digital Storytelling Department
ZGDV e.V.
Fraunhofer Str. 5
64283 Darmstadt
ido.iurgel@zgdv.de

Gary Bente*
*Department of Psychology
University of Cologne
Bernhard-Feilchenfeld-Str. 11
50969 Köln
bente@uni-koeln.de

Abstract

Referring to a recent controversy in psychological literature on facial displays and the expression of emotion it is argued that concerning the implementation of communication abilities in e.g. embodied agents, researchers should stress motivation instead of emotion. Based on a social communicative view it seems feasible to focus on motivation and social intentions when planning the behavior of embodied conversational agents. An architecture implementing emotions and one focussing on motivation (in terms of intentions to affect the user in a certain way) are contrasted. An example of a system based on the latter concept is presented and empirical evidence that this can successfully affect the user is given.

1 Introduction

Embodied conversational agents are expected to on the one hand yield human computer interaction more natural and intuitive and on the other hand to "manipulate" (in a positive sense) the users´ moods and emotions - e.g. by calming them down, elicit positive emotions or alerting the user when necessary. In this paper, it is argued that it is not compulsory to implement emotions and subsequent motivations in order to achieve this ultimate goal to affect the users´ emotions. Instead, the direct implementation of intentions and social goals is proposed as alternative approach. A set of rules derived from this approach is presented that links intentions to specific (nonverbal) behaviors that are known to elicit specific effects in a human observer. Thus, although in humans we can assume that motivation and emotion are highly intertwined and have mutual interaction we propose to not implement this within embodied conversational agents. For agents that merely interact with a social but not with a physical environment the implementation of a goal and corresponding behavior initiation would be more important than to implement evaluation and emotional states. Two studies - conducted with agents whose behavior is controlled based on this approach - show that the direct implementation of social motivations and intentions is effective in leaving favourable impressions and inducing positive feelings in the user.

2 Implementation of emotions

Many architectures of more sophisticated embodied conversational agents incorporate emotions. Emotions are seen as necessary since they yield human-like emotional behaviour that may motivate or affect the user and/or eventually permit empathic behaviour of agents (Picard & Cosier, 1997; Breazeal, 1998; Elliott & Brzezinski, 1998; Elliott, Rickel und Lester, 1999; Picard, 1999; Lester, 2000). Implementations thus are based on assumptions and knowledge derived from emotion research (e.g. Ortony, Clore und Collins, 1988). Internal system states are implemented that relate to basic emotions (e.g. "delight" about correct input) and subsequently motivate a corresponding behavior (e.g. smiling, see e.g. Lester et al., 2000, and their emotive-kinesthetic behavior framework). These approaches are based on psychological theories that presume a direct link between emotion and nonverbal behavior (Tomkins, 1962; Rimé, 1983; Ekman, 1997). On the other hand, of course, there are numerous researchers already focussing on the functions and effects of nonverbal communication regardless of emotive states (mostly in the realm of linguistic or discourse supporting aspects, see Bickmore, 2004; Cassell & Bickmore, 2001; Cassell

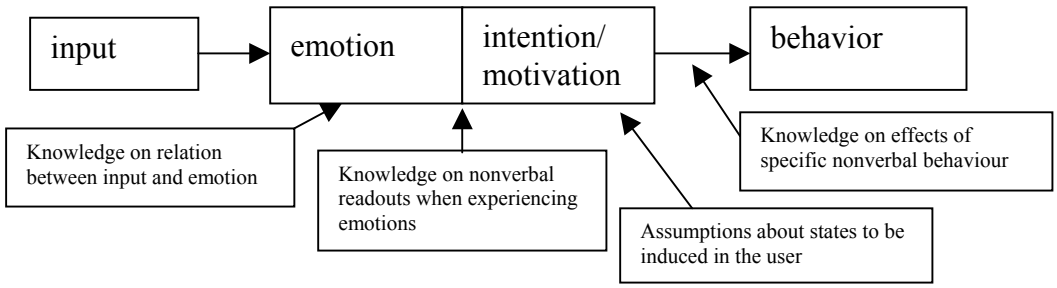


Figure 1: Architecture and necessary knowledge when implementing emotions

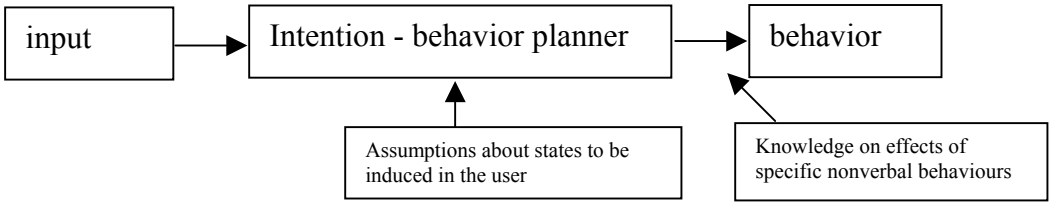


Figure 2: Alternative approach focussing on the state/emotion that is to be elicited in the user

et al., 1999). But as depicted above a considerable number of other researchers still holds the view that implementing emotions and motivation is essential – especially when attempting to influence the human user´s emotion and motivation as e.g. in e-learning scenarios.

In order to implement this "emotion view" one has to possess knowledge about various relations (see figure 1). It has to be known (a) what emotions emerge given a specific input or situation and (b) which nonverbal readouts result given a specific emotion. Finally – since the ultimate goal is to affect the users´ emotion and behavior – one has to know (c) which state should be induced in the user and (d) by means of which cues this can be achieved.

Especially with regard to the first two aspects one has to fall back on knowledge derived from emotion research. But especially the knowlegde concerning the relation of emotions and their behavioural readouts or "expressions" is ambiguous. The assumption that emotion and expression are directly linked and that emotional states involuntarily lead to expressions specific for the respective emotion (Tomkins, 1962; Ekman, 1997) has been challenged in recent years: Researchers of the social-communicative view (Chovil, 1991; Fridlund, 1991) argue that emotional nonverbal behaviors are not determined by emotional states but by social intentions. Referring to empirical findings and evolutionary psychology Fridlund (1991) argues in his behavioral ecology view that it is not functional to directly show ones emotional states but to use ones emotional displays independent from the actual emotional state in a socially reasonable and manipulative way (e.g. not to cry when one is saddest but to cry when assistance is most readily available). In sum, behavior (like e.g. facial displays) is seen to be motivated by social goals and intentions and the concept of emotion is explicitly rejected as not useful when discussing the determinants for facial displays and other "emotive" bahaviors. It is crucial to note that certainly researchers have tried to resolve the debate by stating that both factors are important (see e.g. Hess, Banse & Kappas, 1995). However, it is often overlooked that this statement eventually is equivalent to the emotion view (see e.g. the concept of display rules, Ekman, 1997), whereas Fridlund would explicitly reject this thesis as he more radically follows the view that emotion and facial displays do not necessarily have a connection.

Hence, getting back to embodied agents this raises the question whether it is necessary to implement emotions when their relation to emotive behavior is that debatable. Instead, it could be more effective to directly implement social motivations and intentions and focus on aspects (c) and (d): the intended effects on the user as well as the question by what cues these are achieved.

Unlike e.g. robots who interact with a physical environment, rely on sensoric feedback and thus might need emotional states (in terms of evaluative

aspects of the relation between an agent and its environment) in order to be able to act autonomously (see e.g Dautenhahn & Christaller, 1997; Cañamero, 2003; for a critical discussion of that notion see Sloman, 2004), embodied conversational agents "merely" interact within a social world. Hence, it is arguable whether they need emotions or whether the focus should rather be on the ultimate goal: the motivation to satisfy the users´ needs – which according to Fridlund (1991) should eventually be achievable even for unhedonic entities[1].

3 Alternative approach: Implementation of motivation in terms of social intentions and social goals

With regard to embodied conversational agents the more relevant questions thus are: Which state should be induced in the user and by means of which cues can this be achieved (see figure 2)? It it suggested to motivate behavior according to the desired effects on the user. In consequence, a specific behavior is shown when it – based on the knowledge about its effects – promises to be effective in manipulating the user in a desired way. In case desired state of the user and a nonverbal cue to achieve this state are known, the behavior can be chosen straightforward without the need to arouse an emotion within the agent: The agent e.g. does not have to be sad when an error occurred; he just has to communicate the error in a way that the user does not return the system angrily (e.g. by displaying regret). A corresponding set of rules can be implemented in a behavior planner.

Following this approach, no emotions have to be implemented. Thus, one would not have to answer questions that need to be answered when following the "emotion view" (or else the seemingly resolution of the debate stating that both emotion and social motives are important). These questions would be: a) Which emotion emerges given a specific situation? and b) Which behaviour is shown given a specific emotion? - the latter being not only more difficult to tackle but also more controversely discussed. Instead, one is able to focus on the question that eventually has to be answered in both approaches: By means of which nonverbal cues are agents able to influence the emotions of the user? Hence, merely the motivation and not the emotion aspect is important to be implemented. The general motivation is to affect the users´ mood. More specifically, depending on the situation the intention could be to e.g. calm the user down, to cheer him up or to alert him. Table 1 shows a list of rules that links speech acts (that are chosen according to a specific situation) and corresponding intentions (in terms of goals what to achieve with regard to the users´ feelings) with specific nonverbal behaviors. The latter are assigned to the intentions based on empirical findings (for an overview on socio-emotional effects of nonverbal behavior see Krämer, 2001). It nevertheless has to be mentioned that more research is needed to reveal the effects of nonverbal cues since especially with regard to subtle dynamics the level of knowlegde is rather poor.

By implementing this kind of "knowledge" on the effects of specific cues the agent is enabled to choose that kind of behavior that best promises to influence the user´s feelings and behavior in a desired way. It subsequently has to be implemented (and thus first of all discussed) what the desired states of a user - given a specific situation - are. Should the agent calm the user down when an error occurred or should he draw the user´s anger to himself? And – most importantly – should the agent be one day capable of deciding on this autonomously based on an ability to reason about emotion? (for a model of agents that reason about human emotions see Elliott & Brzezinski, 1998; Gratch & Marsella, 2004). In future, not only this but also various other topics connected to the implementation of motivation and emotion will have to be discussed against the background of both empirical evidence and ethical considerations.

4 Empirical evidence

In order to analyse whether approach and rules are effective, we implemented the rules and empirically veryfied a selection. We conducted two experiments testing whether an embodied conversational agent whose facial displays were based on the rules would influence the users´ feelings more effectively than an agent without facial displays.

The agent was developed by our project partners of the Computer Graphics Center (ZDGV, Darmstadt) within the project EMBASSI (Electronic multimodal service assistance, funded by the German Ministry for Education and Research). Supported by the dialogue architecture developed within the project the agent is able to understand and

[1] Additionally, the implementation of emotions certainly is compulsory when trying to verify a model of human emotions and thus using the embodied agent as a means for fundamental research.

Table 1: Speech acts, corresponding goals and behavior

Speech act (Searle, 1969)	Intention/Goal	Agent behavior	Time
message_greeting: agent greets user	User feels welcomed in a friendly and polite way	Eye brow raise, eyes widen smile head tilt (10 degrees to the right)	start (1 second) before addressing user, after addressing user after addressing user (4 seconds)
message_inform [status: **warning**]	User feels urgency	Eye brow raise, eyes widen Raise hands, palms towards user	During utterance Start of utterance, then drop slowly
message_inform [status: **busy**]	User stays patient	Tilt head slightly downward Look away from user (turn stays with agent)	During utterance After utterance
message_inform [status: **error**] or [status: **failed**]	User is calmed down/appeased	Regret/sorrow display smile head tilt (10 degrees to the right)	During utterance After utterance After utterance (4 seconds)
message_inform [status: **ok**]	Positive feelings are induced	Slight smile	After utterance
message_inform [status: **offer**] (proactive offer)	User is interested	Raise index finger Slight smile	First 3 seconds During utterance, intensify after utterance
message_ reject: (the user´s claim is rejected)	User does not get angry	Neutral display	During utterance
message_command: (proposals or urgent hints)	User feels urgency	Eye brow raise, eyes widen Raise hands, palms towards user	During utterance Start of utterance, then drop slowly
message_cancel: sudden turn-yield after interruption by user	User realizes that he is allowed to speak	Look at user	Until user starts to speak
message_acknowledge: (declares successful execution)	Positive feelings are induced	smile (emblematic gesture)	During utterance, intensify after utterance
query_input: (agent awaits user´s input)	User feels invited to speak	Look at user Slight head tilt (5 degrees to the left)	After utterance 3 seconds after end of utterance (lenght: 5 sec)
query_selection: (list of alternatives)	Comprehension is facilitated	beat gestures (or deictic gesture) Look at user Slight head tilt (5 degrees to the left)	Alongside every alternative After utterance 3 seconds after end of utterance (length: 5 sec)

answer the users´ questions and instructions within the domain of TV and VCR programming. Although the agent allows for free dialogues in the TV/VCR domain, the dialogue within the study was pre-written and only partly interactive with the agents presenting alternatives and the user choosing from them (answers could be typed in via keyboard). In order to meet the criteria of psychological research this approach was chosen to guarantee standardised conditions.

Within the first experiment we compared an agent without, an agent with facial expression and as further control condition a merely audio based system. 60 participants in a between subjects design were randomly assigned to one of the three groups. All participants were asked to complete a task that involved the programming of a VCR. As dependent variables we assessed both behavioral data and subjective self-report data. The participants´ readiness to delegate the task to the agent was assessed by logging the actual behavior of the participants: At one point they were asked whether they chose to do the programming by themselves supported by the agent or whether they preferred to delegate the whole task. The choice was made more relevant than in usual laboratory studies by the fact that participants were told that they would be given an additional incentive when the programming proved to be successful. Additionally, by means of a post-hoc questionnaire we assessed the feelings during interaction (20 items), person perception with regard to the agent (33 items) and evaluation of the system (19 items). Within analyses we wanted to

show whether participants actually are more inclined to delegate the task, to give positive evaluations and whether they felt better when there is an agent (compared to audio). Also, we expected that there would be a difference between the agent showing facial displays derived from the model depicted above and the agent without facial expressions – with the former triggering more delegation, more positive evaluations and feelings.

Table 2: Design of first experiment

Independent Variable	Audio	Agent without facial expressions	Agent with facial expressions
Participants	21	20	19
Dependent Variable	Decision to delegate; Feelings during interaction; Person perception; Evaluation		

Results show that neither the agent with nor the agent without facial expressions led to a higher rate of participants delegating the task - compared to the audio condition. Also, there are no differences concerning participants´ feelings during interaction. With regard to person perception of the agent, though, one out of seven factors that were derived from the 33-item scale showed significant differences: The agent displaying facial expressions was rated as more approachable as the one without and as the audio condition ($F = 7.24$; $df = 57$; $p = .002$).

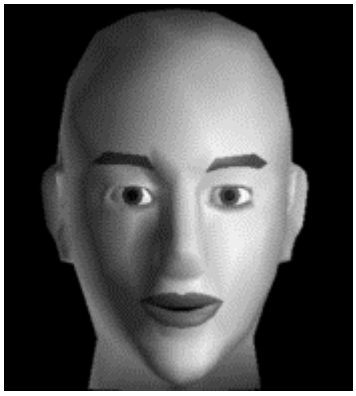


Figure 3: Agent used in experiments

A second study yielded more pronounced differences. Here, 60 participants were infomed by an agent that the VCR programming of a movie they had done previously had failed and the movie had not been recorded. After an additional inquiry whether he may provide another service, the agent says goodbye. Conditions differed with regard to the facial expressions of the agent: One agent did not show any facial expressions, one showed a sorrow/regret display during the first part while smiling again during good bye, one showed a sorrow/regret display during the whole interaction. Here, we wanted to test whether consistent behavior of the agent would lead to more positive evaluations and feelings than both the agent with inconsistent and the agent without facial expressions. Dependent variables again focussed on feelings during interaction as well as on person perception of the agent.

Table 3: Design of second experiment

Independent Variable	Agent without facial expression	Agent with sorrow/regret display in first part	Agent with sorrow/regret display during whole interaction
Participants	18	20	22
Dependent Variable	Feelings during interaction; Person perception; Evaluation		

Concerning feelings during interaction, the agent with inconsistent facial expression receives worst ratings: participants describe themselves as less awake ($F = 3.5$; $df = 57$; $p = .037$) and curious ($F = 4.35$; $df = 57$; $p = .018$) than especially compared to the agent with sorrow/regret display during the whole interaction (see also differences hinting in the same direction but only significant on the 10% significance level: *bored* $F = 2.47$; $df = 57$; $p = .093$; *attentive* $F = 2.67$; $df = 57$; $p = .078$; *committed* $F = 2.98$; $df = 57$; $p = .059$). Also, this is reflected with regard to the factor "naturalness/vitality" ($F = 3.09$; $df = 57$; $p = .053$) that is one of four factors derived from the person perception items: The agent with inconsistent facial displays is rated as least natural. More importantly, concerning the factor "likeability/pleasantness" the agent with consistent sorrow/regret display is evaluated most positive ($F = 4.03$; $df = 57$; $p = .053$). In sum, the agent with a consistent sorrow/regret display leaves a more favourable impression and influences the user in a more positive way than both agent without facial expressions and agent with inconsistent facial display (see also results of Isbister and Nass, 2000, who show that consistency across different communication channels is likewise important).

These results certainly do not prove that implementing merely intentions and a corresponding behaviour planner based on knowledge on the effects of nonverbal cues is a better way to tackle the task of building believable and effective agents. It shows, though, that especially in situations when mood management with regard to the user is needed (e.g. when the user is prone to be disappointed or angry, see above), facial expressions as derived from the rule based behaviour planner can be effective. Thus, the approach is at least worthwhile pursueing.

5 Conclusion

In conclusion we ask: Why not try and build embodied conversational agents that *merely* want – in terms of being motivated to elicit specific states in the user - without implementing emotions (that they would not "feel" anyhow). Not in all cases it seems to be inevitable or necessary to implement human-like attributes. When focussing on the emotions of the user, research efforts would be drawn to the question by which nonverbal cues the users´ emotions can be manipulated most effectively. This could prove to make research more straightforward and some applications easier to realise.

Acknowledgements

This research was supported by the German Ministry of Education and Research (BMB+F) within the project EMBASSI (Multimodal Assistance for Infotainment and Service Infrastructures, BMB+F grant number 01 IL 904 L). We also thank two anonymous reviewers for their suggestions.

References

T. Bickmore. Unspoken rules of spoken interaction. *Communications of the ACM*, 47 (4), 38-44, 2004.

C. Breazeal(Ferrell). Regulating Human-Robot Interaction using `emotions', `drives', and facial expressions. *Proceedings of 1998 Autonomous Agents workshop, Agents in Interaction -- Acquiring Competence Through Imitation*, Minneapolis, MO, 14-21, 1998.

L. D. Cañamero. Designing Emotions for Activity Selection in Autonomous Agents. In R. Trappl, P. Petta, S. Payr (eds.), *Emotions in Humans and Artifacts* (pp. 115-148). Cambridge: The MIT Press, 2003.

J. Cassell, T. Bickmore. Negotiated Collusion: Modeling Social Language and its Relationship Effects in Intelligent Agents. *User Modeling and User Adaptive Interaction 13(1),* 89-132, 2003.

J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, H. Yan. Embodiment in conversational interfaces: Rea. *CHI´99 Conference Proceedings* (pp. 520-527). Association for Computing Machinery, 1999.

N. Chovil. Social determinants of facial displays. *Journal of Nonverbal Behavior, 15* (3), 141-154, 1991.

K. Dautenhahn, T. Christaller. Remembering, rehearsal and empathy - towards a social and embodied cognitive psychology for artefacts. In S. O´Nuallain, P. McKevitt & E. MacAogain (Eds.), *Two Sciences of Mind*. Amsterdam: John Benjamins, 1997.

P. Ekman. Expression or communication about emotion. In N. L. Segal & G. E. Weisfeld (Eds.), *Uniting psychology and biology: Integrative perspectives on human development* (pp. 315-338). Washington: American Psych. Association, 1997.

C. Elliott, J. Brzezinski. Autonomous Agents as Synthetic Characters. *AI Magazine*, 19 (2), 13-30, 1998.

C. Elliott, J. Rickel & J. Lester. Lifelike pedagogical agents and affective computing. An exploratory synthesis. In M. Woolridge & M. Veloso (Eds.), *Artificial intelligence today* (pp. 195-212). Berlin: Springer, 1999.

A. J. Fridlund. Evolution and facial action in reflex, social motive, and paralanguage. *Biological Psychology, 32* (1), 3-100, 1991.

J. Gratch, S. Marsella. A Domain-independent framework for modeling emotion. *Journal of Cognitive Systems Research,* 5 (4), 269-306, 2004.

U. Hess, R. Banse, A. Kappas. The intensity of facial expression is determined by underlying affective state and social situation. *Journal of Personality and Social Psychology, 69* (2), 280-288, 1995.

K. Isbister, C. Nass. Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies, 53*(1), 251-267, 2000.

N. Krämer. *Bewegende Bewegung. Sozio-emotionale Wirkungen nonverbalen Verhaltens und deren experimentelle Untersuchung mittels Computeranimation*. Lengerich: Pabst, 2001.

J. C. Lester, S. G. Towns, C. B. Callaway, J. L. Voerman, P. J. FitzGerald. Deictic and emotive communication in animated pedagogical agents. In J. Cassell, J. Sullivan, S. Prevost & E. Churchill (Eds.), *Embodied Conversational agents* (pp. 123-154). Boston: MIT Press, 2000.

A. Ortony, G. L. Clore, A. Collins. *The cognitive structure of emotion.* New York: Cambrige University Press, 1988.

R. W. Picard, G. Cosier. Affective Intelligence – The missing link. *BT Technology, 14* (4), 150-161, 1997.

R. Picard. *Affective Computing*. Cambridge: MIT Press, 1999.

B. Rimé. Nonverbal communication or nonverbal behavior? Towards a cognitive-motor theory of nonverbal behavior. In W. Doise & S. Moscovici (Eds.), *Current issues in European social* psychology (pp. 85-141). Cambridge: Cambridge University Press, 1983.

J. R. Searle. *Speech acts*. Cambridge: Cambridge University Press, 1969.

A. Sloman. *Do machines, natural or artificial, really need emotions?* Talk at the Cafe Scientifique & Culturel, May 2004. Available: http://www.cs.bham.ac.uk/research/cogaff/talks/#cafe04 [14.1.2005].

S. S. Tomkins. *Affect, imagery, consciousness. The positive affects*. New York: Springer, 1962.

The emotive episode is a composition of anticipatory and reactive evaluations

Mercedes Lahnstein
Department of Electrical and Electronic Engineering
Imperial College London
Exhibition Road, London, SW7 2BT
m.lahnstein@imperial.ac.uk

Abstract

A synthetic research approach to investigating the real time dynamic development of emotive processes is presented, revealing disassociable evaluative component systems leading to "wanting" and "liking" during the emotive episode. A distributed network of motor primitives and forward models is presented, which allows a robot to anticipate the outcome of its actions and to initiate and direct its movement according to self-generated expectancies. The operation of the network is organised according to the signalling properties of the dopaminergic neuromodulatory system, which signals reward according to a prediction error. The prediction error allows to determine the on- and offset of the anticipatory phase, revealing internally encoded expectancies during the emotive process. The computational architecture is tested in four experimental trials, each being comprised of a complete perception-action cycle with internally generated evaluative feedback, allowing the robot to associate the hedonic experience with the action leading to it and to adapt its expectations to changing reward values.

1 Introduction

A rich body of research has focused on the functional role of emotion and motivation in social interaction and communication (Breazeal, 2003; Staller and Petta, 2001), and the role emotion and motivation play in designing action selection and decision-making policies (Avila-Garcia et al., 2003; Bryson, 2003; Canamero, 2003; Cos-Aguilera et al., 2003; Sawada et al., 2004). Work in progress, presented here, takes a synthetic research approach (Pfeifer and Scheier, 1999) using *robot-based experiments* and focuses on the *dynamics* of the *emotive process* itself.

This approach is related to research into value systems in the area of developmental robotics. A number of value systems have been realised in robotic systems, where they are employed to modulate learning and to realize value-dependent modifications for motor activation (Almassy et al., 1998; Krichmar et al., 2000; Pfeifer and Scheier, 1997; Sporns and Aleksander, 2002). The contribution presented here is different, since it emphasizes the developmental nature of the emotive process itself and concentrates on investigating internally encoded expectancies in cognitive and movement dimensions. These expectancies are encoded as associatively learned knowledge about a rewarding stimulus and the movement leading to the rewarding event and are pivotal for initiating and executing goal-directed movements.

Focussing on the real time dynamic development of the emotive episode allows to research the compositional and systemic nature of the process and the reciprocal interaction between different component systems. The three organismic component systems of the emotive episode investigated in this paper are: 1) the cognitive system (predictions, evaluations), 2) the motor system (motor programs), and 3) the sensory monitoring system (feelings), with feelings being the parallel reflection of all ongoing changes in the different component systems during the emotive episode. This approach allows not only to model the phenomenological distinctiveness of the emotive episode during the anticipatory and reactive phase, but also to investigate temporally distinct evaluative components leading to "wanting" and "liking", two seperable systems linking motivation and emotion. Reward information processing involves the ascend-

ing dopaminergic neuromodulatory system of the brain, which is involved in sensorimotor processes that are important for "wanting", i.e. for movement activation and responsiveness to conditioned rewarding stimuli.

This paper proposes that the emotive process is comprised of an anticipatory and a reactive phase, based on qualitatively and quantitatively different and temporally distinct evaluative feedback components on sensory, cognitive and motor levels. The development of the emotive episode is proposed to be multi-factorial and modelled as a function of expected reward value and the actual value of the hedonic experience, with its subsequent decay. A distributed network of motor primitives and forward models is presented, which allows a robot to anticipate the outcome of its actions and to initiate and direct its movement according to self-generated expectancies. Network operation is organised according to the signalling properties of the dopaminergic neuromodulatory system, which processes reward information according to prediction errors (Schulz, 2004, 2002, 2000; Schulz and Dickinson, 2000). The timing of these prediction errors determine the onset and offset of the anticipatory process and reveals internally encoded expectancies.

Drawing upon the neuroscience, psychology, and epigenetic robotics literature, a biologically inspired perspective is taken to investigate the coding of value during the anticipatory phase of the emotive process, aiming to illustrate the dynamical and reciprocal link between motor behaviour, expectancies and goals in general. In particular, the research addresses the suggestion made by Holland and Gallagher (2004) that 'estimates of expected value of the consequences of actions might be frequently intertwined with expectancy'. Results substantiate this suggestion and propose this process to be based on the circular self-organising interaction between expectancies and motor programs, and to be reliant primarily on competition, mutual inhibition and quality-of-prediction-based selection of a winner. This proposed mechanism may be used to inform neurobiological research into orbitofrontal-amygdala interaction.

2 Architecture

A distributed network of *motor primitives* and *forward models* (Demiris and Johnson, 2003) is implemented, which allows a robot to anticipate the outcome of its actions and to initiate and direct its movement according to self-generated expectancies.

The core feature of this active anticipatory system is the pairing of multiple inverse models operating in parallel with corresponding forward models to create a perception-expectancy-execution sequence. Network operation is organised according to the dopamine prediction errors, which determine the onset and offset of the *expectancy process*. Evaluative feedback generated by the prediction errors provides the system with particular expectancies that guide perception and limit action, consistent with cognitive accounts of the role of anticipation in the perception-action cycle.

The computational architecture solves the problem of initiating goal-directed, intentional movement by using motor primitive programs in a dual and parallel control loop. Forward models provide a) the predicted potential outcome of each simulated motor primitive (*motor primitive expectancies*) towards achieving the goal, as well as b) the predicted results of those actions. The architecture chooses the action with the highest expectancy by internally generating it and a winner-take-all approach is used to determine which motor primitive is used at each time step towards achieving the goal (Demiris and Johnson, 2003). This prediction-comparison process achieves the matching between cognitive and motor information. Simultaneously, the system performs the execution of these motor behaviours.

2.1 State representation and generation

The system's state space is represented as a vector of values, comprising interoceptive, i.e. proprioceptive and hedonic states, and exteroceptive states, i.e. quantities derived from the relation of the system to objects in its environment, such as distance.

The dopamine prediction error at a particular time is computed as the difference between the reward value experienced at that time and the reward value predicted for that time:

$$\begin{aligned} Dopamine\ Response_t = \\ Actual\ Reward\ Value_t - \\ Predicted\ Reward\ Value_{t-1} \end{aligned} \quad (1)$$

Learning involves minimising the prediction errors and modifies both performance and prediction according to the prediction error (Pagnoni et al., 2002; Schulz, 2004; Schulz and Dickinson, 2000;

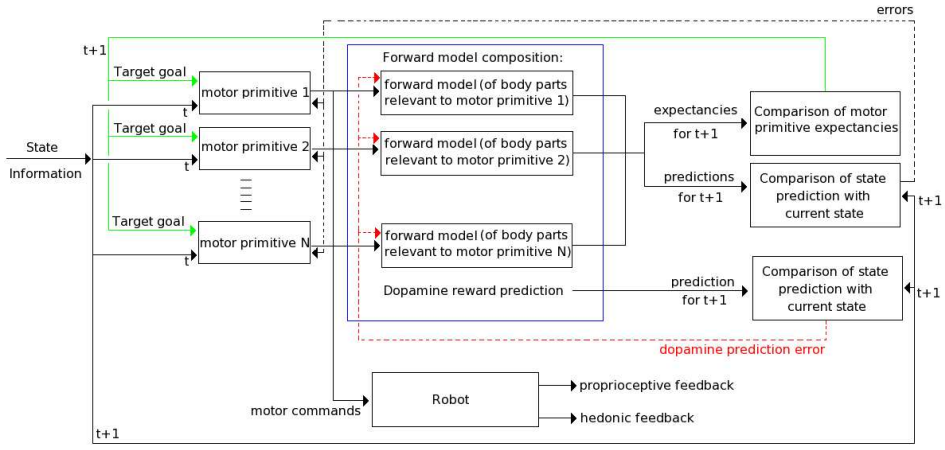


Figure 1: A top view of the computational architecture

Schulz, 1998; Schulz et al., 1997). After learning, the reward value is fully predicted and the error term is zero. By signalling rewards according to a prediction error, dopamine responses provide a teaching signal for model learning, show adaptive responses during associative learning episodes, and transfer from the actual rewarding event to the earliest reward-predicting event, the latter evoking an expectation of the reward value, marking the evaluation leading to "wanting" and initiating the approach behaviour. Accordingly, visual detection of a reward-predicting object results in the generation of a dopamine prediction error, which marks the onset of the anticipatory phase. The expectation to receive a reward and the motor primitive expectancies are generated from the time of object detection and specified in value by the forward model predictions. The hedonic experience of reward marks the offset of the anticipatory phase and only leads to the generation of a second prediction error if the expected reward value does not equal the actual one.

The development of the emotive episode is multi-factorial and modelled here as a function of expected reward value, motor primitive expectancies and the actual value of the hedonic experience, with its subsequent decay. Consistent with this conceptualisation, the emotive episode is comprised of an anticipatory and a reactive phase. The onset and offset of the anticipatory phase is determined by the dopamine prediction errors, and the reactive phase starts at the time of the actual rewarding experience and ends with this hedonic experience having decayed over time.

The prediction for the expected reward value during the anticipatory phase determines the positive valence (hope, optimism, positive expectancy) of the developing emotive episode. The episode undergoes a phase change at the time of the actual rewarding event, when the actual reward value is compared to the predicted value. This evaluation leads to feelings of happiness, satisfaction and contentment if the expected value equals the actual value; leads to increased intensity of feeling happy if the actual reward is higher than the predicted reward value; and to feeling disappointed if the actual reward value is smaller than the predicted value (Buck, 1999).

Due to the lack of existing biological data, a 1:1 mapping between the magnitude of the reward signal and the value of the hedonic experience is chosen. This assumption allows us to focus the investigation on the effects of the cognitive evaluation of expected and actual reward value on the subjective experience - the "liking" component. This sensory component of the emotive episode is chosen to decay exponentially according to:

$$\begin{aligned} Hedonic\ Experience_t = \\ Reward\ magnitude \times e^{-kt}, \\ with\ the\ decay\ rate\ k = \frac{x^{-1}}{2} \end{aligned} \tag{2}$$

Since the decay is dependent upon knowing the elapsed time after the reward signal, the variable x

is introduced in order store this information and to derive the decay value at any given point in time after the reward signal has been given. This allows the modelling of the decay of the emotive episode so that the higher the actual reward value, the slower the decay of the hedonic experience, or alternatively, the smaller the hedonic experience, the faster the decay.

2.2 Implementation of Forward Models

Forward models can be used to not only simulate musculoskeletal aspects but also cognitive and affective system components, thus allowing the prediction of multiple future system states. Each forward model receives as inputs the current states, the motor command generated by the inverse model it is coupled to, and has access to associative memory. Forward models are implemented as integrators to generate one-time-step predictions for guidance of movements.

During the anticipatory phase, forward models generate two kinds of predictions: **a)** motor command hypotheses, which are generated on the basis of the current distance, i.e. predictions are made of what the distance to the object is expected to be if this particular motor command was to be executed. These predictions are compared every cycle to evaluate the motor primitive, which contributes most towards goal completion. The expectancy of each motor primitive is calculated according to:

$$Expectancy = associated\,reward\,value \times normalised\left(\frac{1}{distance}\right) \quad (3)$$

Predictive evaluative feedback generated this way specifies how to initiate and guide movement. Thus, this feedback not only encapsulates the motivation to reduce the distance towards achieving the goal, but also the means to achieve this. **b)** In parallel, predictions are generated for interoceptive and exteroceptive robot states. These predicted states are fed into the inverse model and are considered when the control signals are generated.

2.3 Implementation of Inverse Models

The inverse model is comprised of six motor primitives: rotate left, rotate right, move forward and open the gripper, move backward, close gripper and stand still. During the anticipatory phase, the generation of motor commands by the inverse model is driven by the motor primitive expectancies. Thus expectancies produce the goals for the motor primitives, so that they generate a motor command that will, at the next time-step, result in a larger expectation to achieve the rewarding experience. Thus, in a circular and interactive manner, do generated motor commands affect the subsequent expectation level and motor primitive expectancies control the execution of movement.

3 Experiments

3.1 Experimental setting

The computational architecture was tested in four experimental trials, each being comprised of a complete perception-action cycle with internally generated evaluative feedback, allowing the association of the object colour and the experienced reward value with the action leading to it. Experiments were performed using a robotics platform with the aim of a) showing approach and grasping behaviours according to self-generated expectancies, b) assessing how well the system adapted its expectations due to changing reward values, and c) demonstrating the effects of changing expectancies on the emotive episode.

The robot used is an ActivMedia Peoplebot with an onboard 800 MHz Pentium III. It is equipped with a Canon VCC4 pan-tilt zoom (PTZ) camera, which was used as the main tracking and range-finding sensor. The ActivMedia Colour Tracker (ACTS) was used for segmentation of the object colour, thus achieving recognition. The software was written in C++ and the artificial equivalent of a reward sensor was implemented to model the functional effect of reward on the emotive episode. All processing was done in real time with one full iteration of the main loop executing in 0.1 seconds.

At the beginning of each experimental trial, the robot faced the table at about 1 meter distance. A blue object was then placed on the table initiating the experiment; and upon grasping the object a single reward signal magnitude was given, with a value chosen by the experimenter. This single signal magnitude was chosen deliberately in order to emphasize the effect of reward evaluation on the subjective experience at this point in time. After the emotive episode had decayed, the robot was reset and a new experimental trial started. During the experiment the robot moved

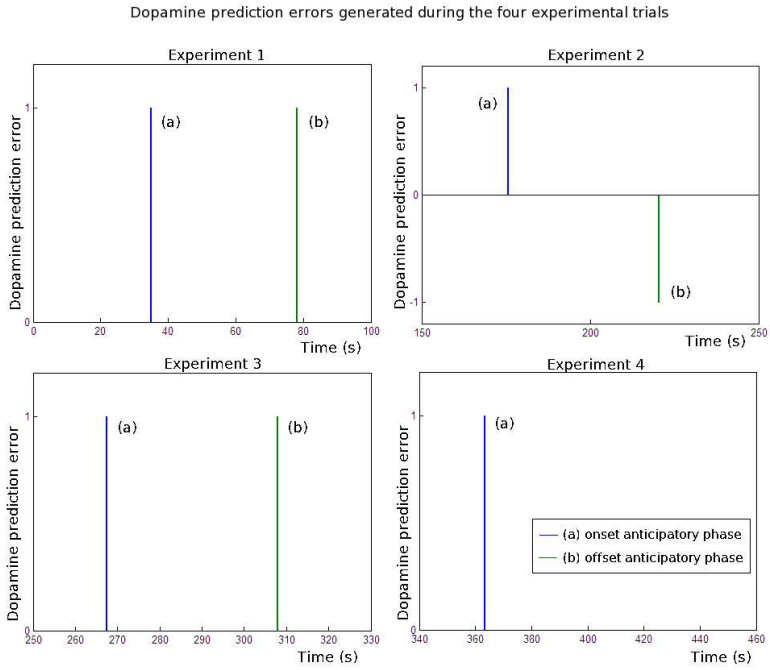


Figure 2: The graphs illustrate different dopamine signalling properties during the learning period over the four experimental trials. Dopamine prediction errors (a) indicate the visual detection of the reward-predicting object. Dopamine prediction errors (b) signal the discrepancy between predicted and actual reward value.

at constant speed.

3.2 Experimental results

The first prediction error in each trial encapsulates the expected reward value learned associatively and the second prediction error marks the difference between the predicted and the actual reward value (figure 2). This second prediction error is positive if the actual reward value is better than expected, negative if it is worse than predicted; and zero if the actual reward value was experienced as predicted.

Results show that the system correctly adapted its expectations to the changing reward values according to the generated prediction error upon receiving the reward signal in each experimental trial. Thus, the system was able to adapt its reward prediction according to changed environmental contingencies prior to initiating the approach behaviour (figure 3). During these experiments inverse model learning was not taken into consideration.

Experimental data plotted in figure 4 demonstrate the generation of motor primitive expectancies, which initiate and direct the robot's movements towards achieving the goal of grasping the object. Re-

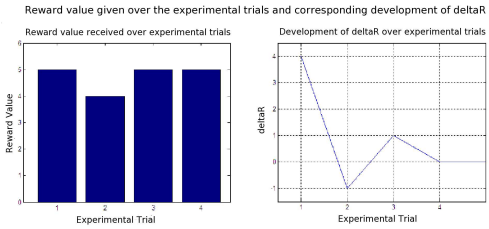


Figure 3: The graph on the left shows the reward values given to the robot in each experimental trial. The graph on the right plots the development of deltaR during learning and the offset of the learning period in experiment 4. The error term deltaR is computed according to equation (1) and is zero after the actual reward value is fully predicted.

sults obtained during the experiments allow visualisation of the developmental nature of the emotive episode and the analysis of feedback generated by sensory, cognitive and motor components during the process (figure 5). These three feedback components were computed according to the equations presented. The expected reward value is encoded in the dopamine prediction error (equation 1), motor primitive expectancies are computed according to equation

3, and the hedonic experience on the sensory level decays according to equation 2.

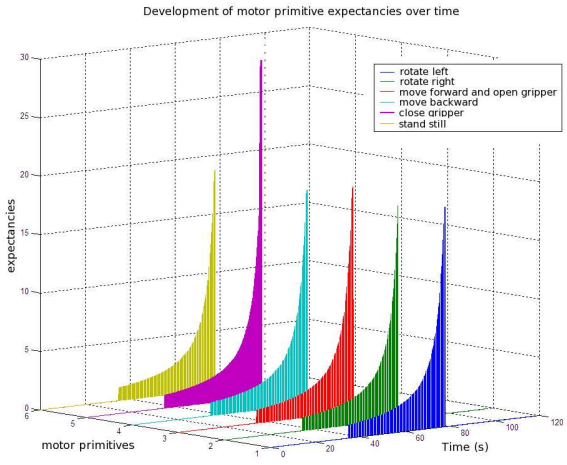


Figure 4: The graph shows the predicted potential outcome of each simulated motor primitive towards achieving the goal. The motor primitives are competing and a winner-take-all approach is chosen to determine which motor primitive is used at each time step.

4 Discussion

The dopamine prediction errors generated during the four experimental trials correctly showed particular signalling properties of this neuromodulatory system. After having learned to associate the hedonic effect with the object colour and the motor primitive 'close gripper', the reward prediction error transferred to the earlier event, the time of visually detecting the object. The association triggered the prediction of the reward value, revealing the internal expectation, and initiated the approach behaviour. Results illustrate the two evaluative and functionally distinct component systems involved in reward information processing. The "wanting" component, which puts motion into emotion and necessitates that the expected rewarding properties of the stimulus must be evaluated and updated prior to initiating approach behaviour, can be disassociated from the "liking" component. The dopaminergic neuromodulatory system is involved in integrating sensorimotor processes that are important for "wanting", i.e. responsiveness to conditioned rewarding stimuli and movement initialisation. Results show that the hedonic assessment of rewards - "liking" - does not depend on dopaminergic neurons, but necessitates the involvement of neural network structures such as orbitofrontal cortex, prefrontal cortex, anterior cingulate cortex and striatum. These structures do not emit a global reward prediction error signal similar to dopamine neurons, but process reward information as transient responses and process the specific nature of the rewarding event. Furthermore, the hedonic assessment of rewards involves opiod mechanisms in the brain stem, nucleus accumbens and pallidum, and manipulation of the opioid system to increase opioid function modulates hedonia.

The robotic system correctly learned to associate the object colour with the hedonic experience and the motor primitive leading to it and adapted its expectations to the changing reward values given over the experimental trials. Thus, the architecture was able to integrate learning and control in a concurrent rather than separate fashion (Pfeifer and Scheier, 1997).

Results show the generation of evaluative feedback on sensory, cognitive and motor levels at different time steps of the emotive process. Associatively learned predictive evaluative feedback on the cognitive level was shown to be encoded in the dopamine prediction errors; to structure the on - and offset of the anticipatory phase; and to drive the expectancy process. Furthermore, results show that the operation of the distributed model of motor primitives and forward models self-organised according to these prediction errors. During the anticipatory phase, this predictive evaluative feedback was continuously generated for movement execution and control via the generation of motor primitive hypotheses about the potential contribution of motor primitives towards the rewarding experience.

The evolution of motor primitive expectancies over time towards achieving the goal of grasping the object clearly demonstrates their development through online interaction with the environment. Results show expectancies to be generated by the close coupling of inverse and forward models, which supports the suggestion made by Holland and Gallagher (2004) that guidance of action is based on expectancies, especially on estimates of the expected value of the consequences of actions. Results show that the implemented distributed network of motor primitives and forward models provides such a mechanism for initiating and guiding movements. The computational architecture uses motor primitive programs in a dual and parallel control loop, which allows generated motor commands

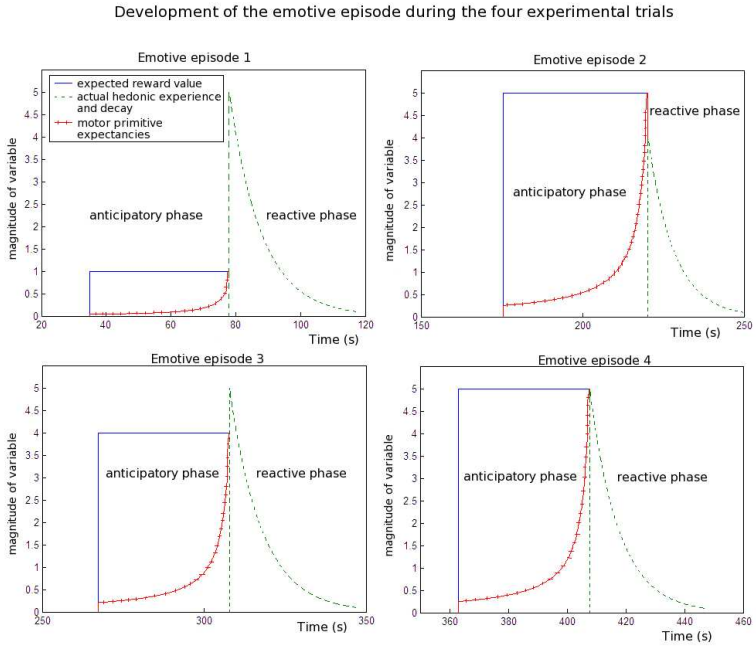


Figure 5: The graphs show the different development of each emotive episode, consisting of an anticipatory (blue and red) and a reactive (green) phase, in each experimental trial. The emotive episode is modelled as function of expected reward value, motor primitive expectancies and hedonic experience. During the anticipatory phase the emotive episode is driven by the predicted reward value (blue) and the motor primitive expectancies (red). After the actual reward signal was given, the hedonic experience and thus the emotive episode decays exponentially. In episode 1, the comparison between predicted and actual value leads to increased feelings of happiness (4 value units). In episode 2, the evaluation leads to feelings of disappointment (1 unit). In episode 3, the result of the cognitive comparison leads to increased feelings of happiness (1 unit). In episode 4, the expected reward value matches the actual one and induces feelings of happiness, satisfaction and contentment.

to affect the subsequent expectation level and in parallel motor primitive expectancies to direct the execution of movement. This circular interaction between expectancies and movement is proposed as a mechanism, which relies primarily on competition, mutual inhibition and quality-of-prediction-based selection of a winner (Demiris and Johnson, 2003).

The bottom-up synthetic approach allowed the modelling of the development of the emotive episode, and reveals that the episode is comprised of an anticipatory and a reactive phase, with evaluations taking place at different time steps leading to phase changes of the episode. The anticipatory phase of the emotive episode starts with the visual detection of the object, is driven by the expected reward value and motor primitive expectancies, and ends with the actual event, at which the expected and the actual hedonic experience are compared. This evaluation leading to feelings of happiness if the actual reward experience is better than predicted or to feelings of disappointment, if the outcome is worse than predicted. Results show that the predictive evaluative feedback generated by the distributed network serves to establish a perception-expectancy-execution sequence, demonstrating the integration of sensory, cognitive and motor feedback and parallel gating of learning during the emotive episode.

Acknowledgements

I would like to thank my supervisors Prof. Igor Aleksander and Dr. Yiannis Demiris for their continuous support and I am grateful to Matthew Johnson and Adam Rae for their support during the software design stage.

References

N. Almassy, G.M. Edelman, and O. Sporns. Behavioural constraints in the development of neu-

ronal properties: A cortical model embedded in a real world device. *Cerebral Cortex*, 8:346–361, 1998.

O. Avila-Garcia, O. Canamero, and R. teBoekhoerst. Analyzing the performance of 'winner-take-all' and 'voting-based' action selection policies within the two-resource problem. In W. Banzhaf, T. Christaller, P. Dittrich, J.T. Kim, and J. Ziegler, editors, *Advances in Arti cial Life: 7th European Conference ECAL*, pages 733–742, Berlin, Heidelberg, 2003. Springer Verlag.

C. Breazeal. Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, 59(1-2):119–155, 2003.

J. Bryson. Action selection and individuation in agent based modelling. In David L. Sallach and Charles Macal, editors, *Proceedings of Agent 2003: Challenges of Social Simulation (to be published)*, 2003.

R. Buck. The biological affects: A typology. *Psychological Review*, 106(2):301–336, 1999.

L.D. Canamero. Designing emotions for activity selection in autonomous agents. In R. Trappl, P. Petta, and S. Payr, editors, *Emotions in Humans and Artifacts*, pages 115–148, Cambridge, MA, 2003. MIT Press.

I. Cos-Aguilera, L. Canamero, and G. Hayes. Motivation-driven learning of object affordances: First experiments using a simulated khepera robot. In F. Detjer, D. Doauterner, and H. Schaub, editors, *The Logic of Cognitive Systems: Proc. 5th Intl. Conference on Cognitive Modeling (ICCM'03)*, pages 57–62, Bamberg, Germany, April 10-14, 2003. Universitats Verlag Bamberg.

Y. Demiris and M. Johnson. Distributed, prediction perception of actions: a biologically insired architecture for imitation and learning. *Connection Science*, 15(4):231–243, 2003.

C. Holland and M. Gallagher. Amygdala-frontal interactions and reward expectancy. *Current Opinion in Neurobiology*, 14:148–155, 2004.

J.L. Krichmar, J.A. Snook, G.M. Edelman, and O. Sporns. Experience-dependent perceptual categorization in a behaving real-world device. In J.A. Meyer, A. Bertholz, D. Floreano, H. Roitblat, and S.W. Wilson, editors, *Animals to Animats 6: Proceedings of the 6th Intl. Conference on the Simulation of Adaptive Behavior*, pages 41–50, 2000.

G. Pagnoni, C.F. Zink, P.R. Montague, and G.S. Berns. Activity in the human ventral striatum locked to errors of reward prediction. *Nat Neuroscience*, 5:97–98, 2002.

R. Pfeifer and C. Scheier. Sensory-motor coordination: The metaphor and beyond. *Robotics and Autonomous Systems*, 20:157–178, 1997.

R. Pfeifer and C. Scheier. *Understanding Intelligence*. MIT Press, Cambridge, MA, 1999.

T. Sawada, T. Takagi, and M. Fujita. Behaviour selection and motion modulation in emotionally grounded architecture for qrio sdr-4 xii. In *Proceedings of 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sendai, Japan, September 28 - October 2, 2004.

W. Schulz. Predictive reward signal of dopamine neurons. *J.Neurophysiol.*, 80:1–27, 1998.

W. Schulz. Multiple reward signals in the brain. *Nat Rev Neurosci.*, 1:199–207, 2000.

W. Schulz. Getting formal with dopamine and reward. *Neuron*, 36:241–263, 2002.

W. Schulz. Neural coding of basic reward terms of animal learning theory, game theory, microeconomics and behavioural ecology. *Current Opinion in Neurobiology*, 14:139–147, 2004.

W. Schulz, P. Dayan, and R.R. Montague. A neural substrate of prediction and reward. *Science*, 275: 1593–1599, 1997.

W. Schulz and A. Dickinson. Neuronal coding of prediction errors. *Annu Rev Neurosci*, 23:473–500, 2000.

O. Sporns and W. Aleksander. Neuromodulation and plasticity in an autonomous robot. *Neural Networks*, 15:761–774, 2002.

A. Staller and P. Petta. Introducing emotions into the computational study of social norms: A first evaluation. *Journal of Arti cial Societies and Social Simulation*, 4, 2001.

Motives inside out

Kamalini Martin
Dept.of Electrical Sciences, Karunya Institute of Technology and Sciences,
Karunyanagar, Coimbatore 641 114 India
kamartin@karunya.ac.in

Abstract

We hypothesize that autonomous agents are systems that proactively respond to input situations or problems by generating, adapting and managing appropriate goals and implementing suitable actions in the context of limited resources. The problem is solved by the right (set of) response(s)/ actions.. Therefore the motivation which drives the generation, adaptation and management of goals and goal directed *actions is or should be derived from the problem that confronts the agent.* In current human self improvement practices, what amounts to a reverse claim is made i.e., that *it is the internal attitude of the human* to a given problem that is more important and not the problem itself. Thus the solution or the response to a problem is generated or at least governed by internal states and is comparatively less determined by the problem. Some philosophers go to extent of stating that the problems themselves are the result of internal attitudes. In this paper, I postulate that under some circumstances, the internal motivation of the agent could shape the external situation, not vice versa, and the interplay between the two amounts to a power struggle.

1, Introduction

The roles and mutual interactions of motivation and emotion in influencing different aspects of cognition and action in artificial agents that interact with their physical environment are to be investigated. In the context of this paper, I define motivation as a derivative of desire, a drive which is to be translated into a description of *what* is to be done, and emotion as a derivative of belief, a state that influences, constrains and directs all functional aspects of cognition in terms of *how* the inputs as viewed, processed and acted on. The problems in this paper are simplistic situations where *common sense* responses are expected. Lack of resources is not considered.

2. Problem Definition

With the given background, the *problem* is described within a situation i.e., an environment in which events or sequences of events arise. The *response* is the perception-cognition-action tuple by which the embodied agent endeavors to solve the problem by perceiving an event and acting on it. The action of the agent directly acts on an event with the intention of altering the situation which caused it, so that problem solving is the main motivation. However the action must cause some change in the situation, hopefully, improving it. This means that situation $\rightarrow$perception $\rightarrow$cognition $\rightarrow$action $\rightarrow$ situation(new) is a loop. Here the word loop is used since the direction from situation to action appears one way, and the feedback from action to perception is through the new situation. Therefore there appears to be a possibility that cognition which dictates action could also drive the situation. This is described by the figure 1.

It is also widely accepted that cognition can modify or filter inputs to perception, so that the same situation can be *viewed* differently by different internal states. The 'new' term in the figure is *virtual* environment – the more common word is 'internal' just as 'external'

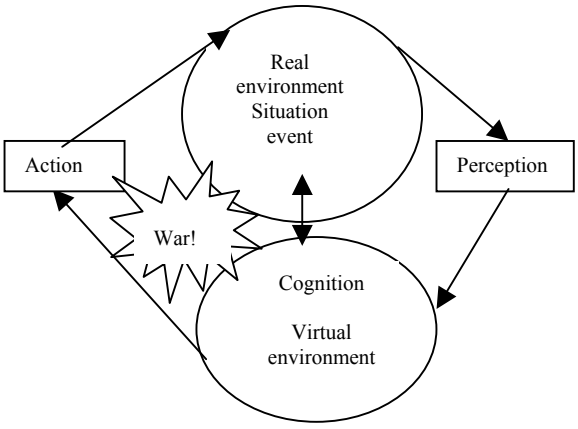


Fig.1 Two conflicting environments

is used more freely than 'real'. The virtual environment could be entirely fictitious, bearing no

relation to the real one. (It could also be an accurate replica in some simple cases). There are many reasons for this is – the real world is too complex to handle, some simplification is required so that important facts are thrown out; the real world is 'viewed' wrongly i.e., attention is diverted, only partial inputs arrive or a wrong filter is imposed on perception; the real world is interpreted wrongly i.e., the translation from perception to cognition is flawed; finally the cognition itself is faulty due to the wrong states and goals of the virtual environment.

When the virtual environment is unlike the real one, there is an interplay which amounts to a power struggle: which one is the key driving force and why?

3. Possible Analysis

The key question is : does the real environment dictate certain actions so that problem solving lies in assessing the situation and acting by rules although within constraints and with resource limitations? Or does the virtual internal environment i.e., the states and processes of cognition governing perceptions and actions, change the real environment? Note that if the virtual environment is very much distorted from the real one, the problem would be worsened by inappropriate actions. This could also be caused by a wrong drive (wrong motivation). In the latter case, since we do not wish to allow the external environment to dominate, problem solving methods would be to monitor and adapt the virtual internal environment and desire, to generate goals that could produce actions that make improvements in the real external world.

Starting arbitrarily from the usual point with the agent already in a state which is either helpful or harmful to the problem solving process

1. The real environment produces an event that is indicative of its situation.
2. The event is translated by (and filtered by internal state) perception into an input for the cognitive process.
3. The input may be a faithful replica of the event, and therefore a representation of the problem or it may be distorted. The extent of distortion lies *purely* in the internal environmental states. This is analogous to saying that the situation is understood or misunderstood solely by the 'mind' and not by the characteristics of the situation itself.
4. Whether the representation is distorted or not, the cognitive process generates actions to deal with it, again, as a function of its own internal states, goals and the filtered (perhaps wrongly) input from perception .
5. The actions may help or harm the situation – this shapes or adapts the external situation to the internal, virtual representation, thus the real environment is modified by the virtual one and not vice versa.
6. Possibly the key point is: who will win? This neither conflict resolution where conflicting systems have a common toplevel goal nor a game of peers with a *similar* goal to win, but opposite reactions. Here, the agent's goal could change in response to the situation. The agent must have *top level* goals, strategies, and resources that can survive all opposing situations.

Conclusion

The 'power' of the 'mind' can be evaluated in two ways – it's 'understanding' i.e., it's ability to produce a representation of the problem accurately (both by filtering of perceptions and actions and its own internal drives and processes) and the effectiveness of the actions it produces. The effectiveness is often a function of the resources it possesses, but not always. Obviously, the more 'powerful' the 'mind' is, the more effect it has on the world it deals with.

Acknowledgents

The author thanks the reviewers for their suggestions, particularly the very appropriate references.

Bibliography

1. A. Ayesh,"Perception and Emotion Based Reasoning: A Connectionist Approach", *Informatica*, vol. 27, pp. 119-126, 2003.
2. P. Carruthers, "The Cognitive functions of language", Behavioural and Brain Sciences (2002), vol. 25, pp.657-726
3. A. Sloman, R. Chrisley and M.Scheutz, "The Architectural Basis of Affective States and Processes" In: *Who needs emotions?: The Brain meets the Machine*, edited by Jean-Marc Fellous and Michael Arbib, 2004.
4. P.Singh and M.Minsky, "An Architecture for combining ways to think", in Proc. Of the Int'l Conf. On Knowledge Intensive Multi-Agent Systems", Cambridge, MA, 2003.
5. M.Minsky, P.Singh and A.Sloman, "The St.Thomas Common Sense Symposium :Designing Architectures for Human-Level Intelligence", AI Magazine, Summer 2004, pp. 113-124.

Synthetic Emotivectors

Carlos Martinho*
*Instituto Superior Técnico
Avenida Professor Cavaco Silva,
2780-990 Porto Salvo, Portugal
carlos.martinho@dei.ist.utl.pt

Ana Paiva†
†Instituto Superior Técnico
Avenida Professor Cavaco Silva,
2780-990 Porto Salvo, Portugal
ana.paiva@inesc.pt

Abstract

This paper presents a simple extension to the base agent architecture supporting the construction of believable synthetic characters: an anticipatory module composed by *emotivectors* that: (1) monitor the information flowing back and forth between the sensor, effector and processing modules of the agent; (2) anticipate the information that will be monitored next; (3) confront the sensor information with the anticipated prediction using a model inspired in Emotion and Attention research that provides the synthetic agent with an autonomous sensation and attention control mechanism aimed at enhancing its believability.

1 Introduction

A critical yet subjective concept to account for when defining the quality of the interaction with a synthetic character is *believability* (Bates, 1994).

Disney animators have been dealing with the creation of believable characters since the dawn of the last century, and have developed a set of guidelines to help in the creation of such believable characters (Thomas and Johnson, 1994). The general principle is to display the internal state of the character to the viewer. This simple principle strives to make the character *aware* of its surrounding environment by reacting emotionally to what happens around it in a consistent way.

The concept of awareness can be further developed into what we call the *behavior loop*: agents should change the focus of their attention and respond emotionally to stimuli provoked by other agents, and these reactions should be responded as well. As an example, if Pluto is laying near the fire when Mickey enters the room, he should react by looking at him and clearly expressing an emotion, perceived as caused by Mickey. In response, Mickey should look back at Pluto and express an emotion. This loop is a simple process that increases the believability of intervening characters.

This work researches which mechanisms are suited to control both the focus of attention and the emotional reactions of a synthetic character, to increase its believability through the behavior loop and strives to make this control as autonomous from the main processing of the agent as possible.

2 Emotivectors

Our synthetic characters are implemented as software agents (Russel and Norvig, 1995). To make control as independent as possible form the agent processing, we extended the architecture with an autonomous module: the *salience module* (see Figure 1).

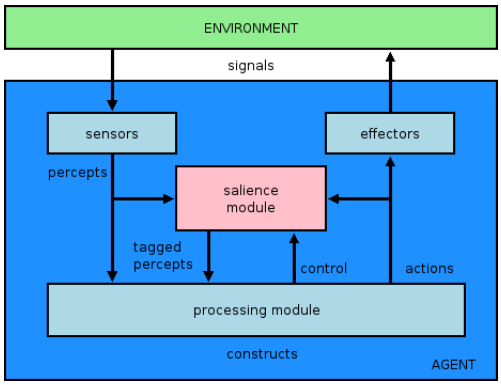


Figure 1: Extended Architecture

The salience module performs a semantic-independent monitoring of the percepts flowing from the sensors to the processing module as well as the action-commands flowing from the processing module to the agent effectors. This monitoring is possible since the code of the information flowing throughout the agent is usually the consistent measurement of a certain aspect of the environment on a same scale over time. Each element of the salience module, responsible to monitor a single piece of information, is called an *emotivector*.

Each emotivector keeps a limited record of its associated signal history, and uses this information to compute a prediction for the next signal value to be read. By confronting the expectation with the actual sensed value, using the affective model described in the next section, the emotivector computes the sensor salience, and adds a tag to the signal containing both attention focus and emotional potential information. The management of all emotivectors according to their salience is performed by the salience module.

3 Affective Model

Rather than striving to implement a detailed affective model (Picard, 1997), we selected a small set of principles from the Psychology of Attention and Emotions. Even these principles fail in providing with an accurate description of how Humans act, we argue that they are useful in building simple synthetic models of behavior that perform well in real time, a critical aspect when considering the creation of "autonomous believability". As both Attention and Emotion cannot be considered separately (Wells and Matthews, 1994), we merged theses principles together into one model. The next subsections describe our approach.

3.1 Attention

Posner (1980) showed that directing attention to a valid location facilitates processing, which led him to suggest that "attention can be likened to a spotlight that enhances the efficiency of the detection of events within its beam". Note that attention is not synonymous with looking. Even when there is no time to make voluntary eye movement to the cued location, facilitation is found. Thus, it seems, attention is oriented to a stimulus.

Posner experimented with central and peripheral cues and found that the attentional spotlight could be summoned by either cues, but peripheral cues could not be ignored whereas central cues could. Posner proposed two attentional systems: an endogenous system, controlled voluntarily by the subject and an exogenous system, outside of the subject control, which automatically shifts attention according to environmental stimuli and cannot be ignored.

After performing a series of experiments clarifying Posner's hypothesis, Muller and Rabbit (1989) verified that exogenous orienting could sometimes be modified by voluntary control, and suggested that "reflexive orienting is triggered and proceeds automatically, and if both reflexive and voluntary orienting mechanisms are pulling in the same direction, they have an additive effect. However, if they are pulling in different directions, their effects are subtractive", which is compatible with Posner's proposal.

Following Posner's proposal, our model uses two interacting components: an exogenous component and an endogenous component. The exogenous component is based on the estimation error and reflects the principle that the least expected is more likely to attract attention. The endogenous component is computed whenever a search value is given to the emotivector: it is a function of the change in the distance to the search value. Unlike the exogenous component that is always positive, the endogenous component is valenced: an increase in the search distance is negative while a reduction is positive.

Both added exogenous and endogenous components define the salience of the emotivector, following Muller's hypothesis. However, an emotivector with a search value also possesses a certain qualia. This is described in the next section.

3.2 Emotion

As Harlow and Stagner (1933), we differentiate between sensation and emotion. Harlow and Stagner proposed that there are basic sensations, innate and undifferentiated, and that emotions are a conditionned form of these sensations, which we learn to refer in particular ways: we are born with the capacity to feel but have to learn the different emotions. Following Harlow's and Stagner's proposal, we assume that emotions are conditioned responses of primary sensations, and concentrate our model in the generation of these sensations. Emotions per se are left to the processing module cognitive or symbolic affective processing.

As Young (1961), we assume that affective processes are defined in a continuum, where changes can occur in either positive or negative direction, giving form to four basic sensations: positive increase, positive reduction, negative increase and, negative reduction. As Young, we give to the affective processes a motivational and regulatory role driving, among other things, the subject toward or away from a stimulus.

Inspired by the behavioural synthesis of Hammond (1970), we considered the emotion as a central state of the organism which is generated by stimuli, both known and unknown. Both stimuli relate to the presence or absence of a reward or punishment. We use the emotivector estimation to anticipate the reward or punishment which, when confronted with the actual

value, triggers one of Hammond's four basic sensations (fear, relief, hope and distress) that we translated to Young's sensations.

Inspired by Millenson (1967), we attribute an intensity to each sensation, which value is the emotional salience, allowing a same sensation to vary accross its dimension, and compute its impact in the conditioning of current operational tendencies. We also use Millenson's designations for our sensations, as the symbols are not connoted to an exact word which, by itself, would imply a certain intensity.

In other words, our affective model considers the following five primary sensations:

Surprise (S) when there is no expectation of a reward or punishment due to the absence of a search value to the emotivector.

Positive Increase (S+) , that we relate to Harlow and Stagner's excitement as well as to Hammond's hope and Millenson positive unconditioned stimulus, and associate with a reward stronger than the expectation. If reward is anticipated and the effective reward is stronger than the expected, a S+ sensation is thrown.

Positive Reduction ($+) , that we relate to Harlow and Stagner's discontentment as well as to Hammond's distress and Millenson reduction of a positive unconditioned stimulus provoking rage, and associate with a reward weaker than expected. If reward is anticipated but the effective reward is weaker than the expected, a $+ sensation is thrown.

Negative Increase (S-) , that we relate to Harlow and Stagner's depression as well as to Hammond's fear and Millenson negative unconditioned stimulus provoking anxiety, and associate with a punishment stronger than expected. If punishment is antecipated and the effective punishment is stronger than expected, a S- sensation is thrown.

Negative Reduction ($-) , that we relate to Harlow and Stagner's pleasure as well as to Hammond's relief and Millenson reduction of a negative stimulus, and associate with a punishment weaker than expected. If punishment is anticipated but the effective punishment is weaker than expected, a $- sensation is thrown.

4 Anticipation

The computation of the emotivector salience rely on the capacity of the emotivector to predict its next state. As there is no a-priori knowledge of the world, we followed a simple assumption: that the intensity of a signal will tend to oscillate around a certain value for a while, and then suddenly change to a random totally new value. We used this model to assess the adequacy of our candidate predictors.

Our chosen predictor is inspired in both the two-phases recirculation algorithm (Hinton and McClelland, 1988), a biologically plausible implementation of the backpropagation algorithm, and Kalman filtering (Kalman, 1960). A detailed description of our lightweight predictor can be found in (Martinho, 2004).

5 Experiment

To test our approach and the strategies to manage several emotivectors at once, we set up a small experiment based on the computer game Dungeon Master[1]. The task is to selected a party of 4 champions to control during the game. The selection is performed by going through the possible 40 champions, one by one.

To find a good starting party, a first approach could be to draw a table of all champions and rate them in terms of their potential value for the party according to some weighting scheme based on their attributes and select the ones with the better scores. However, players running through the game for the first time do not spend around two hours annotating all the champions attributes and computing their potential suitability before entering the game per-se!

Imagine yourself walking through the Halls of Champions, examining each candidate, one by one. You would look at a first one, then at a second, and suddenly you would remark: "This champion has a very high strength. He could make a good warrior!". Later, while observing the statistics of another, you would remark: "This one has a very low dexterity, she would never make to be my ranger, and for what I have seen until now, it will be hard to get a good one..." Step by step, your attention will be drawn to one or other attribute, you will react emotionally and, in a matter of minutes, a party will come together. No notes needed, only the memory of one or other value, and that is exactly the human-like behavior we are aiming for.

[1] Dungeon Master ©1987 Software Heaven Inc., edited by FTL Games, one of the first real-time role playing games.

6 Preliminary Results

We associated an emotivector with each attribute rating the champions. Each time a new champion was observed, its attributes were fed to their related emotivectors and they would each produce a salience value. The salience module would then, using a specific strategy, decide which emotivectors would be sent to the agent for processing, based on the individual saliences. If a sensation was present, it would be expressed by the agent.

We evaluated a simple winner-takes-all strategy and the results were promising. After a few runs with different permutations of the champion order, the algorithm found sub-optimum parties which members were rated as top-5 in the exhaustive approach: enough to start the game knowing to have (maybe not the best) but a good starting party.

However, under certain conditions, an emotivector could hide another, and we are starting to evaluate other possible strategies: salience ordering, threshold salience and meta-estimation.

7 Conclusions and Future Work

We presented an extension to the base agent architecture aimed at enhancing the believability of synthetic characters built upon it. We introduced the salience module, that manages emotivectors: independent mechanisms that monitor the different dimensions of the agent perceptions, estimate their next state and, based on a model of attention and emotion, provide with information regarding both the attention focus as well as the sensation potential of the signal. An interesting aspect is that the salience is computed independently from the semantics of the signal and from the rest of the agent processing. A simple experience was described that allowed us to verify the potential of the approach and test a simple management strategy. The most important result, however, is to have shown that it is possible to enhance synthetic characters with human-like behavior using very simple strategies.

Some aspects are still under development: the emotional exogenous control based on Damasio's somatic markers that is being implemented as described in (Martinho et al., 2003) and; although the model proved to be adequate in providing with mechanisms of human-like behavior, the degree of believability gained using this technique still has to be evaluated in more detail.

Acknowledgements

This work is supported by the EU project "MindRACES: from reactive to anticipatory cognitive embodied systems" (IST-511931).

References

J. Bates. The role of emotions in believable agents. Technical report, Carneggie Mellon University, 1994.

L. Hammond. *Physiological Correlates of Emotion*, chapter Conditioned Emotional State. Academic Press, 1970.

H. Harlow and R. Stagner. *Psychology of Feelings and Emotions*, chapter 2 - Theory of Emotions. Number 40. Psychological Reviews, 1933.

G. Hinton and J. McClelland. Learning representations by recirculation. In D. Anderson, editor, *Neural Information Processing Systems*. American Institute of Physics, 1988.

R. Kalman. A new approach to linear filtering and prediction problems. *Transactions of the ASME, Journal of Basic Engineering*, 1960.

C. Martinho. Synthetic emotension. Technical report, Instituto Superior Tecnico, 2004.

C. Martinho, M. Gomes, and Ana Paiva. Synthetic emotension. In *IVA*, pages 57–61, 2003.

J. Millenson. *Principles of Behavior Analysis*. Macmillan, 1967.

H. Muller and P. Rabbit. Reflexive orienting of visual attention: time course of activation and resistance to interruption. *Journal of Experimental Psychology: Human Perception and Performance*, (15), 1989.

R. Picard. *Affective Computing*. MIT Press, 1997.

M. Posner. Orienting of attention. *Quaterly Journal of Experimental Psychology*, (32), 1980.

S. Russel and P. Norvig. *Artificial Intelligence: a Modern Approach*. Prentice Hall, 1995.

F. Thomas and O. Johnson. *The Illusion of Life*. Hyperion Press, 1994.

A. Wells and G. Matthews. *Attention and Emotion - a Clinical Perspective*. Psychology Press, 1994.

P. Young. *Motivation and Emotion*. Wiley, 1961.

Models of misbelief: Integrating motivational and deficit theories of delusions

Ryan McKay[*†]
[*]Department of Neuropsychology
The National Hospital for Neurology and Neurosurgery,
Box 37, Queen Square, London
WC1N 3BG, United Kingdom
ryan.mckay@uclh.org

Robyn Langdon[†]
[†]Macquarie Centre for Cognitive Science
Macquarie University,
Sydney NSW 2109, Australia
robyn@maccs.mq.edu.au

Max Coltheart[†]
[†]Macquarie Centre for Cognitive Science
Macquarie University,
Sydney NSW 2109, Australia
max@maccs.mq.edu.au

Abstract

Humans are agents that want and like, and the impact of our desires and preferences upon our ordinary, everyday beliefs is well-documented (Gilovich, 1991). The influence of such motivational factors on delusions, which are instances of pathological *misbelief*, has tended however to be neglected by certain prevailing models of delusion formation and maintenance (e.g. Ellis and Young, 1990; Stone and Young, 1997; Davies & Coltheart, 2000; Langdon & Coltheart, 2000; Davies, Coltheart, Langdon & Breen, 2001). This paper explores a distinction between two general classes of theoretical explanation for delusions; the *motivational* and the *deficit*. Motivational approaches view delusions as extreme instances of self-deception; as defensive attempts to relieve pain and distress. Deficit approaches, in contrast, view delusions as the consequence of defects in the normal functioning of belief mechanisms, underpinned by neuroanatomical or neurophysiological abnormalities. It is argued that although there are good reasons to be sceptical of motivational theories (particularly in their more floridly psychodynamic manifestations), recent experiments confirm that motives are important causal forces where delusions are concerned. It is therefore concluded that the most comprehensive account of delusions will involve a theoretical unification of both motivational and deficit approaches. An attempt is made to develop just such a rapprochement, taking as its point of departure a current cognitive neuropsychiatric model of delusion formation, the two-deficit model of Coltheart, Langdon, Davies and Breen.

1 What are Delusions?

If illusions involve low-level misperceptions of reality, then delusions involve cases of high-level *misbelief* – instances where the avowed contents of an individual's beliefs run counter to a generally accepted reality. The prevailing diagnostic view of delusions is that they are rationally untenable beliefs that are clung to regardless of counter-evidence and despite the efforts of family, friends and clinicians to dissuade the deluded individual (American Psychiatric Association, 1995).

Delusions are observed in an array of psychiatric and neurological conditions. They have been referred to as "the *sine qua non* of psychosis" (Peters, 2001, p. 193); together with hallucinations, delusions constitute first-rank symptoms of psychotic disorders such as schizophrenia, schizophreniform disorder, schizoaffective disorder and delusional disorder. Such disorders affect around one percent of the population and have devastating consequences in terms of suffering and loss of functioning. Delusions also occur in association with dementia, temporal lobe epilepsy, Huntington's disease, Parkinson's disease, multiple sclerosis and traumatic brain injury.

Delusions can vary both thematically and in degree of circumscription. Thematically speaking, delusions range from the bizarre and exotic (e.g. the delusion that one's head has been replaced by a pumpkin or that one has been raped by the devil) to the *relatively* humdrum (e.g. an unjustified conviction regarding the infidelity of a spouse, or an overwhelming suspicion of persecution by one's neighbours). This is a nosologically important distinction,

as the presence of bizarre delusions satisfies the symptom criteria for a diagnosis of schizophrenia (even in the absence of other psychotic symptoms), while precluding a diagnosis of delusional disorder.

In terms of scope, delusions vary from the circumscribed and monothematic to the widespread and polythematic (Langdon & Coltheart, 2000). A patient with "Capgras" delusion, for example, may believe that a loved one (usually a spouse or close relative) has been replaced by a physically identical impostor, yet remain quite lucid and grounded on other topics. Other individuals evince a more extensive loss of contact with reality. Nobel laureate John Nash, for example, believed not only that aliens were communicating with him, but also that he was the left foot of God and the Emperor of Antarctica (David, 1999).

2 Theoretical Approaches: Motivational versus Deficit

There have been many proposed theoretical explanations of delusions (for interesting reviews see Winters & Neale, 1983; Blaney, 1999; Garety & Freeman, 1999). Among the various models that have been put forward can be discerned two general classes of theoretical explanation, the *motivational* and the *deficit* (Blaney, 1999; see also Winters & Neale, 1983; Hingley, 1992; Bentall, Corcoran, Howard, Blackwood & Kinderman, 2001; Venneri & Shanks, 2004). In brief, theories of the first type view delusions as serving a defensive, palliative function; as representing an attempt (however misguided) to relieve pain, tension and distress. Such theories regard delusions as providing a kind of psychological refuge or spiritual salve, and consider delusions explicable in terms of the emotional benefits they confer. This approach to theorizing about delusions has been prominently exemplified by the psychodynamic tradition with its concept of *defense*, and by the philosophical notion of *self-deception*. From a motivational perspective delusions constitute psychologically dexterous "sleights of mind", deft mental manoeuvres executed for the maintenance of psychic integrity and the reduction of anxiety.

Motivational accounts of delusions can be generally distinguished, as a major explanatory class, from theories that involve the notion of *deficit* or *defect*. Such theories view delusions as the consequence of fundamental cognitive or perceptual abnormalities, ranging from wholesale failures in certain crucial elements of cognitive-perceptual machinery, to milder dysfunctions involving the distorted operation of particular processes. Delusions thus effectively constitute disorders of belief – disruptions or alterations in the normal functioning of belief mechanisms such that individuals come to hold erroneous beliefs with remarkable tenacity.

A deficit approach to theorizing about delusions would seem to be implicit in the field of *cognitive neuropsychiatry* (David & Halligan, 1996). Cognitive neuropsychiatry is a branch of cognitive neuropsych*ology*, a discipline which investigates disordered cognition in order to learn more about normal cognition (Ellis & Young, 1988; Coltheart, 2002). Cognitive neuropsychiatry involves applying the logic of cognitive neuropsychology to psychiatric symptoms such as delusions and hallucinations (Ellis & Young, 1990; Stone & Young, 1997; Langdon & Coltheart, 2000). The aim of cognitive neuropsychiatry is thus to develop a model of the processes underlying the normal functioning of the belief formation system, and to explain delusions in terms of damage to processes implicated in this model of normal functioning.

Perhaps the best way to represent the distinction between the motivational and deficit approaches is to contrast a motivational account of a particular delusion with a deficit account of the same delusion. Let us take as our example the Frégoli delusion, first described in 1927 by Courbon and Fail (see Ellis, Whitley & Luaute, 1994). Patients suffering from the Frégoli delusion believe that they are being followed around by a familiar person (or people) who is in disguise and thus unrecognizable. The delusion was named after an Italian actor renowned for his ability to impersonate people (Ellis & Young, 1990). A motivational explanation of a particular case of this delusion was suggested by Collacot and Napier (1991; cited in Mojtabai, 1994), who argued that a case of Frégoli delusion in which the patient misidentified certain unknown people as her deceased father might be explicable in terms of wish fulfillment. The development of this woman's delusional belief is here viewed as serving a psychological function, namely gratifying her wish that her father still be present. This explicitly motivational formulation, with its notion of wish fulfillment, is exquisitely Freudian, and consistent with a long tradition of psychodynamic theorizing.

Such an account brooks comparison with the deficit explanation of Davies and Coltheart (2000). Davies and Coltheart integrate a key notion from prevalent deficit accounts of the aforementioned Capgras delusion (the belief that a loved one has been replaced by an impostor), which implicate a dissociation between different components of face recognition (e.g. Ellis and Young, 1990; Stone & Young, 1997). The proposal involves two components of face recognition, an overt "pattern-matching" component and an affective component which provides an experience of "familiarity" when we encounter people we

know. Whereas prevailing deficit accounts of the Capgras delusion suggest that it stems from a diminished affective response to familiar faces (see below), Davies and Coltheart (2000) propose that the Frégoli delusion involves a *heightened* affective response to *un*familiar faces. The ensuing discordance between an experience of the way a stranger looks (unfamiliar, unrecognizable) and the way they "feel" (familiar) might lead to the adoption of the Frégoli belief (cf. Ellis & Young, 1990).

3 The Two Deficit Model

The notion that anomalous perceptual experiences may stimulate delusional hypotheses is a key element of a current model of delusion formation and maintenance known as the "two deficit" or "two factor" model (Davies & Coltheart, 2000; Langdon & Coltheart, 2000; Davies et al., 2001; Coltheart, 2002). This model incorporates an *empiricist* perspective on delusion formation (Campbell, 2001), taking as its point of departure theoretical work by Maher and colleagues (e.g. Maher & Ross, 1984). Maher maintained that delusions do not arise via defective reasoning, but rather constitute rational responses to unusual perceptual experiences, which are in turn caused by a spectrum of neuropsychological abnormalities. Coltheart, Davies, Langdon and Breen agree that such anomalous experiences may indeed be necessary for the development of delusions, and they allocate such experiences the status of Deficit-1 in their two-deficit theory.

An experiment conducted by Ellis, Young, Quayle and de Pauw (1997; see also Hirstein & Ramachandran, 1997) provided support for Maher's contention that delusions are responses to anomalous perceptual experiences. Ellis et al. (1997) recorded skin-conductance responses (SCRs – an index of autonomic activity) while showing Capgras patients and control participants a series of predominantly unfamiliar faces, with occasional familiar faces interspersed. They found that whereas control participants showed significantly greater SCRs to familiar faces than unfamiliar faces, Capgras patients failed to demonstrate a pattern of autonomic discrimination between familiar and unfamiliar faces, showing SCRs of equivalent magnitude to photographs of both types.

Further support for the claim that anomalous perceptual experiences are implicated in the formation of delusions comes from the work of Breen, Caine and Coltheart (2001). These authors investigated the rare delusion of mirrored-self misidentification, whereby patients misidentify their own reflected image. Breen et al. thoroughly examined two patients with this delusion, and found that whereas the first patient (FE) demonstrated a marked deficit in face processing, the second patient (TH, whose face processing was intact) appeared to be 'mirror agnosic' (Ramachandran, Altschuler & Hillyer, 1997), in that he showed an impaired appreciation of mirror spatial relations and was unable to interact appropriately with mirrors. These findings implicate two potential routes to development of the mirrored-self misidentification delusion, underpinned by two types of anomalous perceptual experience; on the one hand an anomalous experience of faces, and on the other an anomalous experience of reflected space.

In addition to their suggestions about Frégoli delusion and mirrored-self misidentification delusion, Coltheart and colleagues identify perceptual anomalies that may potentially be involved in a series of other delusions, including delusions of alien control, thought insertion and Cotard delusion (the belief that one's self is dead). These researchers note, however, that such first-deficit experiences are not sufficient for the development of delusions, as some individuals with similar anomalous perceptual experiences do not develop delusory beliefs about those experiences (see, for example, Langdon and Coltheart's [2000] discussion of delusional Capgras patients versus the non-delusional patients with damage to bilateral ventromedial frontal regions of the brain tested by Tranel, Damasio and Damasio, 1995). Coltheart and colleagues thus claim that Maher's account is incomplete, and invoke a second explanatory factor – a deficit in the machinery of belief revision. Individuals with this second deficit, it is hypothesised, are unable to reject implausible candidates for belief once they are suggested by first-factor perceptual anomalies.

4 Backlash

It would appear that the advent and ascent of rigorous cognitive and neurological models of mental disorders has occasioned something of a backlash against historically prevalent psychodynamic modes of theorizing (Gabbard, 1994). In the field of delusions, recent years have seen psychodynamic accounts usurped by their cognitive neuropsychiatric counterparts. Influential cognitive neuropsychiatric accounts such as that of Ellis and Young (1990) and Stone and Young (1997), which explain delusions as the output of a faulty cognitive system, disregard psychodynamic influences in favour of more austere psychological factors (i.e. "cold cognitive" factors). Such authors view psychodynamic approaches as at best inadequate (Stone & Young, 1997), and at

worst "sterile… tired and outdated" (Ellis, 2003, pp. 77-78). Likewise, the two deficit theory of Coltheart and colleagues, outlined above, which aims "to explain delusions of *all* types" (Langdon & Coltheart, 2000, p. 184, italics in original), contains little provision at present for motivational factors.

Psychodynamic theorists and practitioners have been roundly censured for their notoriously unsound methodologies and outrageous theoretical presumption. Cognitive neuropsychiatric accounts (Ellis & Young, 1990; Stone & Young, 1997; Davies et al., 2001), by contrast, are elegant and theoretically rigorous, yielding empirically testable predictions. Cognitive neuropsychiatric research has shown that at least some delusions are neuropsychological in origin. One wonders, therefore, whether all delusions might be adequately explained in terms of neuropsychological damage, in which case motivational ideas could be dispensed with altogether.

5 Persecutory Delusions

The motivational explanation of Frégoli delusion discussed above strikes one, at least initially, as rather fanciful and far-fetched. The manufacture of Frégoli symptoms seems, after all, a rather convoluted route for the psyche to take in order to satisfy a wish for the continued presence of a deceased relative. The fact that this account fails completely as an explanation for cases of Frégoli delusion where strangers are misidentified as known, but hostile, persecutors, poses an additional obstacle to its success.[1]

Claims about motivational causes of delusion are more plausible elsewhere, however, and the domain of paranoid and persecutory beliefs is an example where there are well-worked out motivational interpretations, notably those of Bentall and colleagues (e.g. Bentall & Kaney, 1996; Kinderman & Bentall, 1996, 1997).

How might a deficit model such as the two-deficit theory of Coltheart and colleagues account for cases of persecutory delusions? In line with the model's empiricist perspective on delusion formation (Campbell, 2001), the first requirement is that a credible candidate for Deficit-1 be proposed. In other words, one needs to identify some kind of anomalous perceptual experience that might plausibly suggest a paranoid delusional hypothesis. The second requirement is that this candidate deficit be present in both deluded and *non*-deluded individuals, i.e. there must exist some individuals with parallel perceptual anomalies who do not develop delusional beliefs grounded in those experiences.

Appropriate candidates for Deficit-1 are not difficult to find. For example, claims of an association between deafness and paranoia have been made for many years (e.g. Piker, 1937). The empirical support for this connection is admittedly somewhat equivocal, with some studies reporting evidence of the association (e.g. Zimbardo, Andersen & Kabat, 1981) and others finding little support for it (e.g. Thomas, 1981). Nevertheless, in terms of the above-stated theoretical requirements, the gradual onset of deafness fits the bill rather well. One can at least conceive of how experiences of surrounding voices at lower than expected volume might stimulate the delusional hypothesis that "people are whispering about me". If coupled with a deficit in belief evaluation abilities (Deficit-2), this dubious hypothesis may be uncritically accepted. If Deficit-2 is not present, the delusional hypothesis will, instead, be rejected and the more plausible belief that one is suffering hearing loss will be adopted.

As noted above, the two-deficit explanatory model of Coltheart and colleagues is intended to encompass all forms of delusional psychopathology, yet makes little provision for motivational causes of delusion. In an attempt to examine the scope of this theory, we have recently conducted a series of empirical investigations of such putative motivational causes, focussing on persecutory delusions. Evidence that motivational factors *do* play a role in the aetiology of persecutory delusions would call for a theoretical overhaul of the two-deficit model in order to incorporate these factors.

5.1 Investigating discrepancies between overt and covert self-esteem

Bentall and colleagues (e.g. Bentall & Kaney, 1996; Kinderman & Bentall, 1996, 1997) are influential advocates of a motivational model of persecutory delusions. Consistent with the traditional psychodynamic emphasis on projection as a mechanism of defence against intolerable inner feelings (Freud, 1895), Bentall and colleagues have claimed that persecutory delusions are constructed defensively, for the maintenance of self-esteem. A key prediction of their model is that persecutory delusions will be associated with a discrepancy between relatively high measures of overt self-esteem and relatively low measures of covert self-esteem. A variety of

[1] This, of course, is not to suggest that all cases of Frégoli delusion will have the same explanation. It is virtually an axiom of cognitive neuropsychology that for many particular symptoms a number of idiosyncratic aetiological pathways are possible (Coltheart, 2002). Breen et al. (2001), for example, identified two potential cognitive neuropsychological routes to development of the mirrored-self misidentification delusion.

studies (e.g. Kinderman, 1994; Lyon, Kaney & Bentall, 1994) have attempted to investigate this hypothesis (for reviews see Garety & Freeman, 1999; Bentall et al., 2001). The findings of such studies, however, are disconcertingly equivocal, with a number of studies suffering from methodological flaws.

We (McKay, Langdon & Coltheart, submitted) have recently examined this hypothesis by utilizing a new and highly influential methodology for eliciting covert effects, the Implicit Association Test (IAT; Greenwald, McGhee & Schwartz, 1998). Following Greenwald and Farnham (2000), our study adapted the IAT for the measurement of covert self-esteem by assessing automatic associations of the self with positive or negative affective valence. Persecutory deluded patients were found to have lower covert self-esteem than healthy controls and remitted patients. On two measures of overt self-esteem, however, the persecutory deluded group did not differ significantly from the other groups once the effects of co-morbid depression had been taken into account. These results are thus consistent with Bentall and colleagues' suggestion that persecutory delusions are associated with a discrepancy between overt and covert self-esteem, and are consistent with psychodynamic accounts of paranoia and persecutory delusions dating back to Freud (1895).

5.2 Need for closure

A second investigation (McKay, Langdon & Coltheart, in preparation) aimed to replicate reported connections between persecutory delusions and need for closure. Need for closure (Kruglanski, 1989; Webster & Kruglanski, 1994) is a motivational construct, associated with a preference for certainty and predictability. Colbert and Peters (2002) have suggested that a high need for closure may account for the tendency of certain individuals with anomalous perceptual experiences to develop delusory beliefs about those experiences. Bentall and Swarbrick (2003) had found that patients with persecutory delusions (both current and remitted) displayed a greater need for closure than healthy control participants. Our study showed that patients with current persecutory delusions scored higher on need for closure than the remitted patients and healthy controls, thus confirming the relationship between persecutory delusions and need for closure.

The above investigations have found compelling evidence that motivational factors play a vital role in the genesis of persecutory delusions. In particular, these studies have shown that persecutory delusions are associated with discrepancies between overt and covert measures of self-esteem, consistent with the defensive theoretical scheme of Bentall and colleagues; and that persecutory delusions are associated with the motivational construct of need for closure. The implication of these findings is that motivational factors *are* important, despite their almost wilful neglect by certain cognitive neuropsychiatric models. How then are we to best theoretically integrate such factors into existing cognitive neuropsychiatric accounts?

6 A Theoretical Synthesis

We argue that although there are good reasons to be sceptical of psychoanalytic theories, empirical studies such as those reported above demonstrate that these theories do contain a notion of key importance for models of delusions and belief formation - the insight that motives can be important doxastic forces (*doxastic* = of or relating to belief). We propose therefore that motives be incorporated into the two-factor scheme of Coltheart and colleagues as a first-factor source of unreliable doxastic input – a means by which individuals prone to the *second* factor are led astray when forming beliefs, such that resulting beliefs track desires rather than reality. In this modified two-factor account of delusion formation, the first factor constitutes whatever sources of information suggest a particular delusory belief, be they anomalous perceptual experiences or defensive desires. Individuals with the "second factor" would be prone to giving undue weight to unreliable sensory information, and liable to having their belief-formation systems derailed and overridden by motivational factors.

How might this motivationally modified two-factor account be applied to persecutory delusions? We have already touched upon the possibility that deafness might constitute a perceptually anomalous Deficit-1 in such delusions. It may be that in certain cases persecutory delusions arise in the context of multiple relevant first-factor sources, including both aberrant perceptual experiences and defensive desires. A man with encroaching deafness, for example, might be highly motivated to avoid any evidence of this infirmity. He might therefore experience an organically underpinned perceptual anomaly – the voices of others at lower than normal volume – in a context of wanting to believe that his faculties are still intact. These two sources of doxastic input – the hearing loss and the desire to deny the hearing loss – might jointly suggest the paranoid belief that others are whispering about him. Such a belief would both account for the perceptual evidence and simultaneously satisfy the desire. Given an additional context of inadequate belief evaluation abili-

ties (Deficit-2), an implausible paranoid hypothesis might be elaborated into a full-blown persecutory delusion rather than being rejected.

The evidence that persecutory delusions are associated with discrepancies between overt and covert self-esteem allows the motivational element in the above story to be further elucidated. The individual described may be motivated to avoid any evidence of his hearing impairment at least in part because such evidence is a threat to his self-esteem. It may be that he has some covert awareness that his anomalous perceptual experiences signal encroaching hearing loss (equating to lowered covert self-esteem). Projection would then constitute the process whereby the perceptual experience and low covert self-esteem suggest the delusional hypothesis: the cause of the perceptual anomaly is projected onto others (externally attributed – see Kinderman & Bentall, 1997[2]), resulting in the belief that others are whispering about him.

Alternatively (or perhaps additionally), motives might enter the story at the level of the second factor, playing a role in the evaluation of doxastic input by constituting constraints on the processing of belief-related information. Westen (1998) has discussed the connectionist notion of *constraint satisfaction*, noting that "Psychodynamic theory can augment a connectionist model in proposing that affects and affectively charged motives provide a second set of constraints, distinct from strictly cognitive or informational ones, that influence the outcomes of parallel constraint-satisfaction processes" (p. 359). Perhaps incoming doxastic information is ordinarily processed so as to satisfy motivational constraints as well as constraints of verisimilitude, in which case it may be a feature of the second factor that the belief-formation system becomes unduly biased toward satisfaction of the former. Need for closure is one such motivational constraint, separate from the alethic injunction to approximate reality. Our individual who experiences being surrounded by low volume voices might be highly motivated to achieve some closure, to account for his anomalous perceptions. That other people are whispering around him might come very readily to mind and, rather than exploring alternative hypotheses, the paranoid belief that others are whispering may provide a satisfactory solution to his immediate alethic and motivational constraints.

Hypotheses such as these seem amenable to investigation via computational modelling techniques. Sahdra and Thagard (2003) have recently applied a computational model of emotional coherence to successfully simulate a case of self-deception taken from Hawthorne's novel *The Scarlet Letter*. These authors expanded an implementation of the theory of explanatory coherence (see Thagard, 2000) by allowing units representing propositions in an artificial neural network to have valences as well as activations. The resulting system successfully self-deceived in that it yielded acceptance of false propositions, consistent with implemented preferences. Comparable simulations might be used to model the beliefs of our hypothetical hearing-impaired individual. For example, to simulate a case of factor one in the absence of factor two, one might utilise an explanatory coherence implementation, such that input is simply the evidential propositions (e.g. "the voices of my colleagues are at lower than normal volume") and the coherence associations between them. To simulate the conjunction of both factors, on the other hand, one might assign valences to units representing propositions (e.g. a negative valence to the proposition "I am deaf"). Such speculations denote an area ripe for future research.

7 Conclusion

Baars (2000) argues that the scientific scepticism regarding psychodynamics is disproportionate, and marvels that "few academic scientists are inclined to simply separate the wheat from the chaff in Freudian thought" (p. 13). We have argued that sound reasons for scepticism notwithstanding, psychoanalytic theories do indeed contain a notion that models of delusions may ignore at their peril, namely the insight that motives can be potent doxastic forces. Taking as our point of departure the two deficit model of Coltheart and colleagues, we have explored a theoretical integration between the motivational and deficit approaches to delusions, with the aim of showing that a single overarching theory is not only scientifically desirable, but theoretically viable.

References

American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, International Version (DSM-IV)*. Washington, DC: American Psychiatric Association, 1995.

[2] Although few studies have attempted to characterise persecutory delusions or persecutory attributions in terms of neuroimaging (Blackwood, Howard, Bentall & Murray, 2001), Blackwood et al. (2000) have found that *depressive* attributions (which, in contrast to persecutory attributions, involve internalising negative events and externalising positive events; see Seligman, Abramson, Semmel & von Baeyer, 1979) require activation of the left precentral gyrus.

B. J. Baars. Conscious emotional feelings – beyond the four taboos: An introductory comment. *Consciousness & Emotion,* 1(1):11-14, 2000.

R. P. Bentall, R. Corcoran, R. Howard, N. Blackwood & P. Kinderman. Persecutory delusions: A review and theoretical integration. *Clinical Psychology Review,* 21(8):1143-1192, 2001.

R. P. Bentall & S. Kaney. Abnormalities of self-representation and persecutory delusions: A test of a cognitive model of paranoia. *Psychological Medicine,* 26:1231-1237, 1996.

R. Bentall & R. Swarbrick. The best laid schemas of paranoid patients: Autonomy, sociotropy and need for closure. *Psychology & Psychotherapy: Theory, Research & Practice,* 76(2):163-171, 2003.

N. J. Blackwood, R. J. Howard, R. P. Bentall & R. M. Murray. Cognitive neuropsychiatric models of persecutory delusions. *Am J Psychiatry,* 158(4):527-539, 2001.

N. J. Blackwood, R. J. Howard, D. H. ffytche, A. Simmons, R. P. Bentall & R. M. Murray. Imaging attentional and attributional bias: An fMRI approach to the paranoid delusion. *Psychological Medicine,* 30:873-883, 2000.

P. H. Blaney. Paranoid conditions. In T. Millon & P. H. Blaney (Eds.), *Oxford textbook of psychopathology* (Vol. 4, pp. 339-361). New York: Oxford University Press, 1999.

N. Breen, D. Caine & M. Coltheart. Mirrored-self misidentification: Two cases of focal-onset dementia. *Neurocase,* 7:239-254, 2001.

J. Campbell. Rationality, meaning, and the analysis of delusion. *Philosophy, Psychiatry, & Psychology,* 8:89-100, 2001.

S. M. Colbert, & E. R. Peters. Need for closure and jumping-to-conclusions in delusion-prone individuals. *Journal of Nervous & Mental Disease,* 190(1):27-31, 2002.

M. Coltheart. Cognitive neuropsychology. In H. *Pashler and J. Wixted (Eds.) Stevens' handbook of experimental psychology (3rd ed.), Vol. 4: Methodology in experimental psychology* (pp. 139-174). New York: John Wiley & Sons, 2002.

A. S. David. On the impossibility of defining delusions. *Philosophy, Psychiatry, & Psychology,* 6(1):17-20, 1999.

A. S. David & P. W. Halligan. Editorial. *Cognitive Neuropsychiatry, 1*:1-3, 1996.

M. Davies & M. Coltheart. Introduction: Pathologies of belief. In M. Coltheart & M. Davies (Eds.), *Pathologies of belief* (pp. 1-46). Malden, MA: Blackwell Publishers, 2000.

M. Davies, M. Coltheart, R. Langdon & N. Breen. Monothematic delusions: Towards a two-factor account. *Philosophy, Psychiatry, & Psychology*, 8(2-3):133-158, 2001.

H. D. Ellis. Book review: Uncommon psychiatric syndromes. *Cognitive Neuropsychiatry,* 8(1):77-79, 2003.

H. D. Ellis, J. Whitley & J-P. Luaute. Delusional misidentification: The three original papers on the Capgras, Fregoli and intermetamorphosis delusions. *History of Psychiatry*, 5(17):117-146, 1994.

A. W. Ellis & A. W. Young. *Human Cognitive Neuropsychology*. Hove, E. Sussex: Lawrence Erlbaum Associates, 1988.

H. D. Ellis & A. W. Young. Accounting for delusional misidentifications. *British Journal of Psychiatry,* 157:239-248, 1990.

H. D. Ellis, A. W. Young, A. H. Quayle & K. de Pauw. Reduced autonomic responses to faces in Capgras delusion. *Proceeding of the Royal Society of London: Biological Sciences,* B264:1085-1092, 1997.

S. Freud. Draft H. In J. Strachey (Ed.), *The Standard Edition of the Complete Psychological Works of Sigmund Freud* (Vol. 1, pp. 206-213). London: Hogarth Press and the Institute of Psychoanalysis, 1895.

G. O. Gabbard. *Psychodynamic psychiatry in clinical practice: The DSM-IV edition.* Washington: APA Press, 1994.

P. A. Garety & D. Freeman. Cognitive approaches to delusions: A critical review of theories and evidence. *British Journal of Clinical Psychology,* 38:113-154, 1999.

T. Gilovich. *How we know what isn't so: The fallibility of human reason in everyday life*. New York: The Free Press, 1991.

A. G. Greenwald & S. D. Farnham. Using the Implicit Association Test to measure self-esteem and self-concept. *Journal of Personality & Social Psychology,* 79(6):1022-1038, 2000.

A. G. Greenwald, D. E. McGhee & J. L. K. Schwartz. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality & Social Psychology,* 74(6):1464-1480, 1998.

S. M. Hingley. Psychological theories of delusional thinking: In search of integration. *British Journal of Medical Psychology,* 65:347-356, 1992.

W. S. Hirstein & V. S. Ramachandran. Capgras syndrome: A novel probe for understanding the neural representation of the identity and familiarity of persons. *Proc R Soc Lond B,* 264:437-444, 1997.

P. Kinderman. Attentional bias, persecutory delusions and the self-concept. *British Journal of Medical Psychology,* 67(1):53-66, 1994.

P. Kinderman & R. P. Bentall. Self-discrepancies and persecutory delusions: Evidence for a model of paranoid ideation. *Journal of Abnormal Psychology*, 105(1):106-113, 1996.

P. Kinderman, & R. P. Bentall. Causal attributions in paranoia and depression: Internal, personal, and situational attributions for negative events. *Journal of Abnormal Psychology,* 106(2):341-345, 1997.

A. W. Kruglanski. *Lay epistemics and human knowledge: Cognitive and motivational bases*. New York: Plenum, 1989.

R. Langdon & M. Coltheart. The cognitive neuropsychology of delusions. *Mind & Language,* 15(1):183-216, 2000.

H. M. Lyon, S. Kaney & R. P. Bentall. The defensive function of persecutory delusions: Evidence from attribution tasks. *British Journal of Psychiatry,* 164(5):637-646, 1994.

B. A. Maher & J. A. Ross. Delusions. In H. E. Adams & P. B. Sutker (Eds.), *Comprehensive handbook of psychopathology*. New York: Plenum Press, 1984.

R. McKay, R. Langdon & M. Coltheart. Jumping to delusions? Paranoia, probabilistic reasoning and need for closure. Manuscript in preparation.

R. McKay, R. Langdon & M. Coltheart. The defensive function of persecutory delusions: An investigation using the Implicit Association Test. *Cognitive Neuropsychiatry,* submitted manuscript.

R. Mojtabai. Fregoli syndrome. *Australian & New Zealand Journal of Psychiatry,* 28:458-462, 1994.

E. Peters. Are delusions on a continuum? The case of religious and delusional beliefs. In I. Clarke (Ed.) *Psychosis and spirituality: Exploring the new frontier* (pp. 191-207). London, England: Whurr Publishers, 2001.

P. Piker. Psychologic aspects of deafness. *Laryngoscope,* 47:499-507, 1937.

V. S. Ramachandran, E. L. Altschuler & S. Hillyer. Mirror Agnosia. *Proc R Soc Lond B,* 264:645-647, 1997.

B. Sahdra & P. Thagard. Self-deception and emotional coherence. *Minds and Machines,* 13:213-231, 2003.

M. E. Seligman, L. Y. Abramson, A. Semmel & C. von Baeyer. Depressive attributional style. *Journal of Abnormal Psychology*, 88(3):242-247, 1979.

T. Stone & A. W. Young. Delusions and brain injury: The philosophy and psychology of belief. *Mind and Language, 12*:327-364, 1997.

P. Thagard. *Coherence in thought and action*. Cambridge, MA: MIT Press, 2000.

A. J. Thomas. Acquired deafness and mental health. *British Journal of Medical Psychology,* 54(3):219-229, 1981.

D. Tranel, H. Damasio & A. R. Damasio. Double dissociation between overt and covert face recognition. *Journal of Cognitive Neuroscience,* 7:425-432, 1995.

A. Venneri & M. F. Shanks. Belief and awareness: reflections on a case of persistent anosognosia. *Neuropsychologia,* 42:230-238, 2004.

D. M. Webster, & A. W. Kruglanski. Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology,* 67:1049-1062, 1994.

D. Westen. The scientific legacy of Sigmund Freud: Toward a psychodynamically informed psychological science. *Psychological Bulletin*, 124(3):333-371, 1998.

K. C. Winters & J. M. Neale. Delusions and delusional thinking in psychotics: A review of the literature. *Clinical Psychology Review*, 3:227-253, 1983.

P. G. Zimbardo, S. M. Andersen & L. G. Kabat. Induced hearing deficit generates experimental paranoia. *Science,* 212(4502):1529-1531, 1981.

Emotions as reasons for action: A two-dimensional model of meta-telic orientations and some empirical findings

Ulrich Mees
University of Oldenburg
D-26111 Oldenburg
Ulrich.Mees@uni-oldenburg.de

Annette Schmitt
University of Oldenburg
D-26111 Oldenburg
Annette.Schmitt@uni-oldenburg.de

Abstract

We are presenting a new model on the relationship between "emotion" and "action". Going beyond existing hedonic concepts of motivation, the model attempts to clarify the influence of emotion on motivation. Here fore, our work contributes to one of the key issues addressed by the Symposium "Agents that Want and Like: Motivational Roots of Cognition and Action".

1 Introduction

Since ancient Greek philosophy, a general hedonic or pleasure principle of motivation has prevailed: People are motivated to approach pleasure and to avoid pain (see f. i. Cabanac, 1992).

From fields such as psychoanalysis to early empirical motivation psychology, this hedonic approach was greeted by many prominent researchers. For instance Atkinson (1957) defined a motive as "a disposition to strive for a certain kind of satisfaction" (p. 36). In continuing these early approaches, Mees and Schmitt (2003) distinguished between *goals* and *reasons* for actions. While a goal specifies what action is taken, the reason for an action answers questions such as *why* or *what* action is taken *for*.

2 A two-dimensional model of meta-telic orientations

We assume the reasons for actions can be found *either* in the hope that certain positive emotions may occur or persist, *or* in the hope that negative emotions can be reduced or avoided. Emotions, or more specifically affective states also including moods, therefore constitute reasons for actions. In more general words: Subjectively, actions have the function either to improve or to retain the affective quality of the actor's experience or, respectively, to avoid or to reduce the worsening of these affective qualities of experience. However, not all actions must correspond *directly* or *immediately* to this purpose.

2.1 Four classes of reasons of action

We have distinguished four classes of reasons why a certain goal is strived for by a person.

a) A person strives for a goal, because the activity the goal pertains to and its outcome are judged positively: A person likes to educate children or enjoys doing sports, etc. The reason for a particular action is the experience of positive, agreeable emotions during or as a direct consequence of this action. Thus, carrying out this action is *directly approach-motivated*, i.e. the person attempts to approach the experience of a desired, agreeable emotion by means of the corresponding action.

b) A person acts in a certain manner in order to reduce an unpleasant emotion. For instance, one does not exercise because it is fun, but rather to watch one's figure and to reduce discontent with a weight considered too high. By doing sports, a person hopes to diminish the disagreeable emotion of discontent. The activity "doing sports" thus is *directly avoidance-motivated.*

c) A person carries out an action because it is regarded as the indirect means to a purpose, which is connected to the experience of positive emotions. For example, a person does sports, because this is regarded as a means to the higher goal "to stay fit and healthy". If this higher goal is reached, the emotion is again pleasant because it increases personal well-being, enhances self-consciousness etc. Again, we are concerned with *approach-motivation*, yet this is an *indirect* form. Actions are carried out in order to achieve something else which results in agreeable affective states.

In everyday life, *why* an action is taken is often answered with higher goals. To answer the question, "Why do you exercise?" it suffices to reply, "Because I want to stay fit and healthy". Adding “And this increases my well-being" is so evident that it seems unnecessary to remark. Nevertheless, the desired experience of these positive emotions is the actual reason for such instrumental actions.

d) Finally, a goal-directed activity can also be carried out in order to avoid offending, hurting or disappointing, etc. significant others. A person carrying out a particular action then avoids unpleasant emotions not directly but indirectly: The action avoids unpleasant emotions in another person. The actor anticipates that if significant others experience negative emotions this will result in negative emotions (such as shame, guilt, fear, sorrow etc.) within oneself. The latter will occur at the latest if the significant other criticizes or reproaches the actor. For example, a person may be motivated to go jogging, because a loved one suggested to do so to improve his or her figure. The motivation to go jogging is *indirect avoidance motivation*: The person does something in order to avoid something else, because this would lead to one's own negative emotions.

Indirect avoidance motivation also refers to cases in which people take, for example, preventative measures against anticipated danger. These also do not directly reduce unpleasant emotions, but rather indirectly avoid them by anticipating their potential occurrence.

2.2 Telic and metatelic orientations

We distinguish between two kinds of action orientation: *telic orientation* (derived from the Greek word telos = goal) describes an individual's intentional orientation towards a certain class of equivalent goals (*what* class of action). By contrast, a *meta-telic orientation* refers to the preference of particular classes of reasons for personal actions (*why* the action is taken; derived from meta- = beyond; meta-telic = what underlies a telic orientation, i.e. its emotional reason).

Our model integrates important distinctions as known in motivation psychology, namely on the one hand the difference between approach and avoidance motivation (see the *regulatory focus theory* by Higgins, 1997), and between intrinsic and extrinsic motivation (see the *self-determination-theory* by Ryan and Deci, 2002) on the other.

2.3 Some empirical findings

We (Mees and Schmitt, 2003) constructed a questionnaire to record fundamental telic and meta-telic orientations and tested it with a sample of N = 267 students. We therefore received respondents' values for the four dispositional meta-telic orientations regarding sixteen telic orientations or sixteen areas, i.e. "partnership", "power and prestige", "competence and curiosity", "children and family", "sports", "aggression and retaliation", "individualism", “health”, “affiliation”, “sex”, “tradition”, “teaching”, “art and culture”, “religiosity”, “prosocial behaviour” and “hedonism”. These sixteen areas have been selected according to Reiss (2000) and are the result of a factor-analysis.

Across all the sixteen areas, the four meta-telic orientations revealed significant positive correlations with each other. This is not astonishing, since every meta-telic orientation deals with reasons for actions and not for their omission.

However, only the two avoidance orientations revealed significant high negative correlations with the neuroticism. Here, the meta-telic orientation for direct avoidance showed a higher positive correlation (r = .37) with neuroticism than meta-telic orientation for indirect avoidance (r = .27) did. In contrast, a significant correlation between the two approach orientations could not be found with the variable "neuroticism".

References

John W. Atkinson. Motivational determinants of risk-taking behavior. *Psychological Review,* 64(6):359-372, 1957.

Michel Cabanac. Pleasure: The common currency. *Journal of Theoretical Psychology*, 155: 173-200, 1992.

E. Tory Higgins. Beyond pleasure and pain. *American Psychologist*, 52(12):1280-1300, 1997

Ulrich Mees and Annette Schmitt. Emotionen sind die Gründe des Handelns (Emotions are the reasons for action). In U. Mees and A. Schmitt (Eds.), *Emotionspsychologie* (Psychology of emotion), pp. 13-101, Oldenburg: BIS, 2003.

Steven Reiss. *Who am I? The 16 basic desires that motivate our actions and define our personalities*. New York: The Berkeley Publishing Group, 2000.

Richard M. Ryan and Edward L. Deci. An overview of self-determination theory: An organismic-dialectical perspective. In E. L. Deci and R. M. Ryan (Eds.), *Handbook of self-determination research,* pp. 3-37. Rochester, NY: The University of Rochester Press, 2002.

Cogito Ergo Ago: Foundations for a Computational Model of Behaviour Change

Cosimo Nobile[*,**]
[*]Department of Computer Science
University of Liverpool
Chadwick Building, Peach Street
Liverpool, L69 7ZF, UK
cosimo@csc.liv.ac.uk

Floriana Grasso[†]
[†]Department of Computer Science
University of Liverpool
Chadwick Building, Peach Street
Liverpool, L69 7ZF, UK
floriana@csc.liv.ac.uk

Abstract

So far AI researches in the health care promotion have considered strategies and techniques for making people aware of their health related problems and helping them to change their behaviour in order to have a better life style and be healthier. Very few researches though, to our knowledge, have focused on the deeper meanings behind a behaviour change. We argue that taking into account cognitive aspects, supported by solid psychological and philosophical theories, might help us to provide the right advice, to the right person, at the right time.

1 Introduction

This research represents our contribution to PIPS, one of the leading projects in the health care delivery arena and funded by the European Union under the FP6[1] Integrated Projects. PIPS, Personalised Information Platform for life and health Services, is a four year project started in January 2004 aiming to improve the current health care delivery models. Recognising the importance of personalised and prevention-focused health care services, PIPS will be providing the right support to the European public by means of special Virtual Agents. These agents will also be in charge of giving health related advice to citizens/patients (helping them to stop smoking, to follow a certain diet, to improve their physical activity etc.). Our own experience and many scientific studies (Prochaska et al. (1995) among others) have proven that changing one's behaviour is not an easy task and, sometimes, represents one of the hardest challenges of our life. Such a change though becomes a critical step to take when its consequences have an impact on our health and our well-being.

We seek to create a computational framework of how changes take place, able to capture and handle the processes behind a behaviour change. Several philosophical, psychological and sociological theories of behaviour change exist and our attempt is to ground our research in some of the most solid theories in those areas. Understanding the hidden causes behind changes will help us not only to model a more believable agent capable of reasoning about, and modifying, its own behaviour, but we also believe it to be essential in order to reason about other agents' motivations and emotional states.

In this paper we present an overview of the theories that seem most appealing to our purposes, then we introduce our proposal for integrating these theories and, finally, we offer some preliminary considerations on computational issues.

2 Theoretical Foundations

Research on Medicine and Nutrition give us details of WHAT needs to be changed in our behaviour in order to have a healthier life-style, prevent or cure diseases and so on. Unfortunately though, this information is not enough for our purposes since they do do not lead us to understand the dynamics behind a behaviour change or, in other words, HOW such a change occurs. The *Stages of Change Model* (Prochaska et al., 1995) defines, instead, very clearly HOW we deal with changing our behaviour by presenting six stages ultimately leading to the change and pointing out the importance of applying different techniques tailored to the particular stage involved. The Stages of Change Model, recognises that behaviour change is a process,

[*]Ph.D. student.

[1]EU 6th Framework Programme in e-Health, FP6/IST No. 507019.

Figure 1: Stages of change

a series of steps (Fig.1)[2], rather than a one-off event.

While giving scientific evidence of this assertion by examining how successful self-changers change, the model identifies stages of change and other factors that predict treatment outcomes. There are six stages of change:

- *Precontemplation*: no intention to change or unaware of the problem.
- *Contemplation*: intention to change but not ready for the action.
- *Preparation*: intention to take action within one month.
- *Action*: behaviour change.
- *Maintenance*: consolidate the result.
- *Termination*: finally out of the problem.

In order to succeed, one must go through all these stages and in the same order (from Precontemplation to Termination). There is always the possibility, though, of returning to some prior stages and this phase is called Relapse. The Stages of Change Model identifies also nine key "change processes" and suggest their use depending upon the particular stage involved. The basic idea is that all individuals have the potential to change. Self-motivated changers are much more effective than guided changers (Prochaska et al., 1995, pg. 21) but the structure of this model can certainly strengthen and significantly improve the chances of succeeding. We strongly believe, though, that the merely knowledge of `WHAT` and `HOW` to change is not enough to create a believable model of behaviour change; we need to have some understanding of the reasons behind our changes or, in other words, `WHY` we change. *Cognitive Dissonance Theory* (Festinger, 1957) proposes the concept of dissonance as one of the main drive in our behaviour. For nearly half a century Festinger's theory has been representing, and still represents, one of the most solid and influential theory in social psychology. Its revolutionary idea is that human mind cannot hold two conflicting thoughts at the same time. It might look a bit too simplistic, but its implications and applications are wide and sometimes unexpected. Most of the smokers, for instance, know that smoking is unhealthy but, careless, they continue to do it. They typically deny the gravity of their habit, or find justifications to smoking, because the alternative would be to face the dissonance between their behaviour and their knowledge. Studies in Health Psychology and Medicine have also demonstrated the existence of relations between various health problems. People who smoke, in fact, are much more likely to develop other bad habits such as poorer diet (Shah et al., 1993), higher alcohol intake (Morabia and Wynder, 1990) and less physical activity, and even ex-smokers can develop bad habits (French et al., 1996). These and many more studies demonstrate how, most of the times, problem behaviours represent only the tip of an iceberg and we believe that Cognitive Dissonance Theory might allow us to understand and fight back all these problems to their very common root. Festinger's theory states that pairs of *cognitions*, that is *"any knowledge, opinion, or belief about the environment, about oneself, or about one's behaviour"*, can be either relevant or irrelevant to one another. Moreover, relevant pairs represent either *consonant* or *dissonant* cognitions. Consonant cognitions occur when they follow from one another, dissonant cognitions occur when the opposite of one of them follows from the other. The size of the dissonance is measured by its magnitude and it is proportional to the importance of the dissonant cognitive elements. This concept has also been extended to groups of cognitive element. The first symptom of dissonance is pressure, a feeling of uncomfortable tension, which can be seen as an attempt of the mind to reduce dissonance (or, at least, to avoid further increases). This Pressure, whose strength is a function of the magnitude of the dissonance (Festinger, 1957, pg. 18), is a very powerful motivator that pushes the individual towards eliminating the dissonance. According to the author, dissonance could be seen as a trigger for a dis-

[2]Source: http://www.cdc.gov/nccdphp/dnpa/physical/starting/ (last updated 6 Feb 2003), Centers for Disease Control and Prevention, Department of Health and Human Services, USA.

sonance reduction's activity as much as hunger triggers a hunger reduction's activity. The tension can be released in different ways:

- by changing dissonant cognitions
- by adding new consonant cognitions
- by reducing the importance of dissonant cognitions

Of course *"The maximum dissonance which can exist between two elements is equal to the resistance to change of the less resistant of the two elements."* (Festinger, 1957, pg. 266).

Dissonance Theory has been generating slightly different variations of the theory itself in the past 50 years. All these revisions, though, have reconfirmed dissonance as a motivation for cognitive changes. Among these, interestingly Aronson (1968) interpreted the theory in terms of the discrepancy between one's *self-image*[3] and behaviour.

A complementary theoretical perspective is given by the concepts of *Reciprocal Determinism* and *Self-Efficacy* (Bandura, 1986). Reciprocal determinism states that a person's behaviour, environment, and psychological processes influence each other in a *"triadic reciprocality"* (Fig.2).

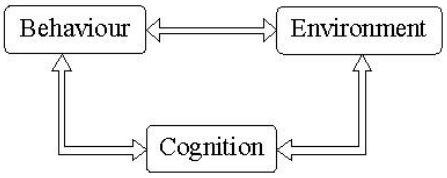


Figure 2: Reciprocal Determinism

Self-efficacy is the *"people's beliefs about their capabilities to produce designated levels of performance that exercise influence over events that affect their lives."* and therefore *"Self-efficacy beliefs determine how people feel, think, motivate themselves and behave."* (Bandura, 1997).

Cognition, in this view, not only plays a critical role in people's capability to adapt, change and self-regulate, but also contribute to create the reality around them. In fact, *"what people think, believe, and feel affects how they behave"* (Bandura, 1986, pg. 25). The problem is that, since we are working with people and not with machines or theorem provers, we must consider that *"people's level of motivation, affective states, and actions are based more on what they believe than on what is objectively true"* (Bandura, 1997, pg. 2).

[3] An individual's conception of himself/herself and his/her own identity, abilities, worth etc.

3 Towards a Cognitive Model of Change

Putting together these views, we can look at behaviour changes from a different perspective: if our self-image determines the way we behave, this means that we could change our behaviour by "simply" changing our self-image. In other words, what we think of ourselves make us act in a certain way. Coherently with the Stages of Change Model, where a bad behaviour cannot be changed whilst still being in the earliest stages of Precontemplation, Contemplation and Preparation, Dissonance Theory explains why it is not possible to jump stages and, even if this happens, why it is not going to last, as the Relapse stage is always on the doorstep, since the self-image is not coherent with the action that has been taken. Dissonance between one's inner and outer self needs to be created and amplified in order to modify his/her behaviour because, as long as one keeps holding an old picture of himself/herself, he/she will simply and coherently behave according to that image.

In conclusion, we interpret each move from one stage to the next one as a cognitive dissonance reduction process. The Stages of Change Model explains very well `HOW` the changes take place, what and in what order the different phases are, whereas the Cognitive Dissonance Theory focuses its attention on the particular individual, on one's self-efficacy, on one's ability of changing the outside by changing the inside first, and move through the stages of change with a new image, from time to time, targeted to the particular processes in each stage.

With this in mind, our efforts are concentrated towards formalising a computational cognitive model of the processes behind a behaviour change. In particular: the *Agent Model* will be a formalisation of the Cognitive Dissonance Theory, specialised to the concept of self-image. The *Change Model* will be instead a formalisation of the Stages of Change Model.

We think of associating different self-images' stereotypes to the different stages in the Stages of Change Model. By making assumptions on what the self-image ought to look like in the next stage the advisory agents will try to help the user in modifying his/her self-image by producing truly tailored advice. We expect our formalisation to be an extension of the classic belief-desire-intention (BDI) architecture.

The feasibility of this approach from the computa-

tional point of view has been reassured by Gawronski and Strack (2004), who observed the *propositional nature* of Cognitive Dissonance Theory.

Moreover, important philosophical studies (Thagard and Verbeurgt, 1998) have proven cognitive dissonance to be essentially a *constraint satisfaction problem*. This idea has led (Shultz and Lepper, 1996) to the formulation of a computational model for cognitive dissonance based on a constraint satisfaction network. This model might be well out of our interests, being a connectionist approach rather than a logical approach, but it still represents a tangible proof of the Cognitive Dissonance theory's computability.

The cognitive theory for *agent communication pragmatic* (Pasquier and Chaib-draa, 2003) applies, instead, the cognitive dissonance theory to multi-agent systems in order to give agent communication more degrees of automation. This computational framework has been successfully employed in modelling dialogue games and simple attitude change processes but, despite being very inspiring, gives only a partial answer to our problem which is modelling behaviour changes in health related domains.

4 Conclusion and Evaluation Issues

In this paper we have presented an overview of some of the most interesting theories behind behaviour change, we have also briefly illustrated our proposal for integrating these theories in a computational cognitive model of change and, finally, we have offered some preliminary thoughts on computational issues.

The work is very preliminary, but, nevertheless, it is progressing by taking advantage of various collaborations with our partners in PIPS and their diversified expertise in health, medicine, nutrition and counselling. They will be providing us assurance about the validity of the theories we refer to and feedback about our results and conclusions. We also plan to evaluate our model against real cases in two different PIPS demonstrators (in Spain and China) before the end of the project in 2008.

Acknowledgements

This work has been supported by the EU under the 6th Framework Programme in e-health, FP6/IST No. 507019.

References

E. Aronson. Dissonance theory: Progress and problems. In *Theories of cognitive consistency: A sourcebook*, pages 5–27. Chicago: Rand McNally, 1968.

A. Bandura. *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice Hall, 1986.

A. Bandura. *Self-efficacy: The exercise of control*. New York: Freeman, 1997.

L. Festinger. *A theory of cognitive dissonance*. Evanston, IL: Row, Peterson, 1957.

S.A. French, D.J. Hennrikus, and R.W. Jeffrey. Smoking status, dietary intake, and physical activity in a sample of working adults. *Health Psychology*, 6(15):448–454, 1996.

B. Gawronski and F. Strack. On the propositional nature of cognitive consistency: Dissonance changes explicit, but not implicit attitudes. *Journal of Experimental Social Psychology*, (40):535–542, 2004.

A. Morabia and E.L. Wynder. Dietary habits of smokers, people who never smoked, and ex-smokers. *American Journals of Clinical Nutrition*, (52):933–937, 1990.

P. Pasquier and B. Chaib-draa. The cognitive approach for agent communication pragmatics. In *Second International Conference on Autonomous Agent and Multi-Agents Systems, AAMAS'03*, pages 544–551, Melbourne, Australia, July 2003.

J. Prochaska, J. Norcross, and C. DiClemente. *Changing for Good*. New York:Avon Books, 1995.

M. Shah, S.A. French, R.W. Jeffrey, P.G. McGovern, J.L. Foster, and H.A. Lando. Correlates of high fat/calorie food intake in a worksite population: The healthy worker project. *Addictive Behaviours*, (18):583–594, 1993.

T. R. Shultz and M. R. Lepper. Cognitive dissonance reduction as constraint satisfaction. *Psychological Review*, (103):219–240, 1996.

P. Thagard and K. Verbeurgt. Coherence as constraint satisfaction. *Cognitive Science*, (22):1–24, 1998.

See What You Want, Believe What You Like: Relevance and Likeability in Belief Formation

Fabio Paglieri
University of Siena
Piazza San Francesco, 53100 Siena, Italy
paglieri@media.unisi.it

Abstract

In this paper, a theoretical analysis and a formal model of the influence of motivations and emotions over belief formation is presented. The basic questions addressed are the following: Whether and how do the agent's goals influence her beliefs? Is this influence a mere bias, or there is a rationale behind it? The inadequacy of current formalisms of belief dynamics in coping with these issues is highlighted, and an alternative model (Data-oriented Belief Revision, DBR) is shortly outlined.

1 Introduction

In spite of overwhelming evidence on the important role played by motivation and emotion in belief formation and change (Festinger, 1957; Kruglanski, 1980; Kunda, 1990; Swann, 1990; Forgas, 1995; 2000; Oatley and Jenkins, 1996; Frijda et al., 2000; Swann et al., 2002), and regardless current interest on affective dynamics in computing and multi-agent systems (Picard, 1997; Cañamero, 2003; Evans et al., 2003), motivational and emotional features still fail to be integrated in most of the formal and computational models of belief dynamics (cf. 2), such as AGM belief revision, Truth-Maintenance Systems, and probabilistic approaches. In addition, all these formalisms suffer from an oversimplified assumption on the status of belief processing: that is, they share the view that belief formation is a straightforward, unique, and highly standardized procedure. In contrast, psychological research (Kruglanski, 1980; Kruglanski and Ajzen, 1983; Kunda, 1990; Forgas, 1995), as well as everyday experience, provides convincing evidence that there are indeed several different ways of developing and assessing our beliefs, which obey to different dynamics and might result in different outcomes, i.e. different sets of beliefs (cf. 3).

Aiming to overcome such limitations[1] of current doxastic formalisms, this paper is focused on the most procedural dynamics of belief formation (i.e. *belief selection*): a formal and computational model of this specific type of belief processing is outlined (cf. 4), in which the effects of motivations and (to a minor extent) emotions are taken into account. It is also argued that motivational and emotional influences at this basic level of belief dynamics have received so far only marginal attention in the literature (for a review, see Kunda, 1990). However, this paper aims to show that motivations and emotions do play an important role also in the most procedural instances of belief formation, namely by determining the *relevance* and the *likeability* of the information being processed by the agent, hence affecting the outcome of her belief selection (cf. 5 and 6). This in turn raises the issue whether similar influences are fully rational features of belief processing, or rather mere distortions that happen to bias human judgement. In the latter case, formal models of rational belief change would obviously be excused for neglecting the role of relevance and likeability. In section 5, I shortly discuss this issue, by pointing out that relevance is indeed a rational feature of human belief dynamics, while likeability requires more cautious assessment.

[1] Similar drawbacks can be held against these formalisms only as far as they are expected to model belief dynamics in cognitive agents. However, their original purposes were often very different (e.g., theory change in the history of science for the AGM model), and only later on extension to belief dynamics in cognitive agents was suggested.

2 Belief Dynamics Without Motivation and Emotion

The most popular formalisms for doxastic dynamics are *AGM belief revision* (Alchourrón et al., 1985; Gärdenfors, 1988; Rott, 2001), *Truth-Maintenance Systems* (Huns and Bridgeland, 1991; Doyle, 1992), and *probabilistic models* (Berger, 1985; Fagin and Halpern, 1994; Boutilier, 1998). In spite of several

significant differences between these approaches (for technical comparison, see Gärdenfors, 1988; Doyle, 1992; Friedman and Halpern, 1999), they all share a disregard for the influence of motivation and emotion over belief formation and change. These formal models consider relevant only *structural properties of beliefs*, i.e. properties depending on internal relations within the belief set: namely, factual credibility in TMS and Bayesian networks (given a certain belief, how many other beliefs support it, and how many counter it?), and epistemic importance in AGM belief revision (how much is central a given belief in the agent's belief set?). While both these criteria are obviously relevant in belief formation, they are not the only ones to play a significant role: to include in the picture motivational and emotional influences as well, a *mapping between beliefs and goals* should be integrated in the model. None of the existing formalisms takes care of that[2]: more generally, motivation is not considered as a driving force in belief dynamics, if not implicitly and in abstract terms – e.g., AGM belief revision can be said to be implicitly driven by the 'motivation' of maintaining a coherent set of beliefs, avoiding contradictions when faced with new information in contrast with previous convictions. But the effect of motivation over belief formation runs much deeper than that, requiring a model capable of handling specific goals and their impact on belief dynamics (cf. 4 and 5).

3 Coming to Believe: Several Paths toward Belief

As a side-effect of neglecting motivation in belief formation, current doxastic formalisms are led to assume that belief formation is a rather simple, unique and standard procedure. However, as soon as we want to move from the field of formal logic and computer science to the domain of cognitive and social psychology, this assumption is shown to be simplistic and misleading – both on the ground of experimental evidence (Kruglanski, 1980; Kruglansk and Ajzen, 1983; Kunda, 1990; Swann, 1990) and concerning theoretical modelling (Forgas, 1995; Miceli and Castelfranchi, 1998; 2000).

[2] This is not exactly true for probabilistic approaches, since the connection between decision-making and beliefs is a well-known topic of interest in Bayesian analyses (Berger, 1985). However, such analyses concern themselves with setting idealized standards of rationality, hence they keep utility (motivation) quite separate from probability (belief). On the contrary, here effects of motivation over beliefs and their perceived strength are discussed, also to show the deep rationale behind such influence (cf. 5).

Indeed, it seems that cognitive agents do apply different procedures and heuristics in assessing their beliefs, depending (among other factors) on their own current motivation (cf. 5). This argues against the adequacy of any generic formalism of doxastic dynamics, and in favour of more differentiated models, aiming to capture specific processes of belief formation within a common framework. Full discussion of these alternative dynamics of belief formation lays beyond the scope of this paper: here only a very broad divide is introduced between more elaborate and deliberate assessment of beliefs (*belief appraisal*), and more procedural and semi-automatic processes of belief formation (*belief selection*).

In partial contrast with most of the literature on motivated reasoning, the following analysis refers specifically to belief selection, focusing on *motivational influences over procedural dynamics in belief assessment.* In fact, one of the aim is to highlight and model *systematic effects of goals over belief formation*, which do not depend on the agent paying explicit attention to the current task – as it happens instead in experimental settings used to test motivated reasoning (Kunda, 1990; cf. 5).

4 Data-oriented Belief Revision

This section shortly outlines a formal and computational framework to model belief selection and change, i.e. *Data-Oriented Belief Revision* (from now on, DBR). DBR was first conceived as an alternative way of modelling belief revision in agent-based social simulation (Paglieri, 2004; Paglieri and Castelfranchi, 2004), trying to overcome several limitations and shortcomings of AGM-style models (Friedman and Halpern, 1999; Segerberg, 1999; Pollock and Gillies, 2000; Wassermann, 2000). Belief formation plays a key role in DBR, since the whole process of belief change is conceived as an *emergent* effect over time of performing data assessment and belief selection on different sets of information (Fig. 1).

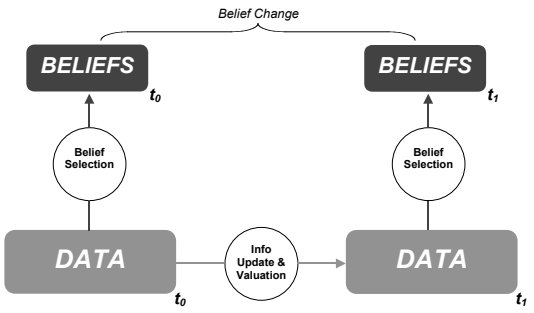


Figure 1: Belief change as an emergent effect

More precisely, in DBR agents can change their belief set either due to corresponding modification in the available data, or because they apply new selection policies over old sets of data: in both cases, belief revision is rooted in belief formation.

4.1 Data and Beliefs

DBR relies on a basic distinction between *data* and *beliefs*. Data are *information gathered by and available to the agent*, stored in her memory as the result of information update (external sources) or inferential reasoning (internal sources); in contrast, beliefs are *data accepted as reliable*, on the basis of a selection procedure characteristic of the agent and based on the informational properties of candidate data (cf. 4.2 and 4.4). This simple distinction, so far ignored (AGM) or marginalized (TMS and Bayesian networks) in formal models of belief change, immediately yields a fairly complex picture of doxastic dynamics (Fig. 2).

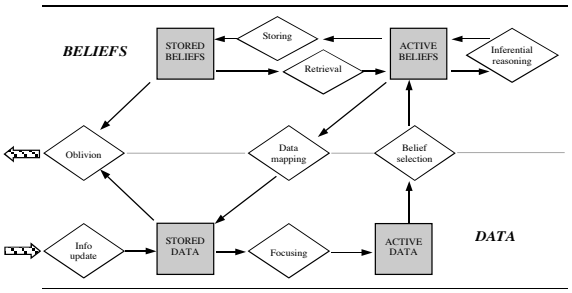


Figure 2: Doxastic processing in DBR

Four properties of data are here introduced and discussed, capturing and refining Castelfranchi's 'reasons to believe' (1996):

I. *Relevance*: a measure of the pragmatic utility of the datum, i.e. the number and values of the (pursued) goals that depends on that datum;
II. *Credibility*: a measure of the number and values of all supporting data, contrasted with all conflicting data;
III. *Importance*: a measure of the epistemic connectivity of the datum, i.e. the number and values of the data that the agent will have to revise, should she revise that single one;
IV. *Likeability*: a measure of the motivational appeal of the datum, i.e. the value of the pursued goal directly fulfilled by that datum.

Finally, this two-layered model of belief dynamics also serves to highlight different functional properties of data and beliefs in DBR, as summarized in Table I. Here the most prominent feature of DBR is the *integration of sub-symbolic and symbolic processing of information* – the former at the level of data, the latter concerning beliefs (a hybrid dynamics quite similar to the one envisioned in the ACT-R architecture: Anderson, 1996; Anderson et al., 2004). While data are structured in networks (cf. 4.3) and their assessment is carried on as a massively distributed process (cf. 4.4), beliefs are organized in ordered sets and processed in a standard sequential fashion, e.g. by selectively applying inference rules over them.

Table I: Data and beliefs: an overview

	DATA	BELIEFS
Basic properties	*Relevance, credibility, importance, likeability*	*Strength*
Organization	*Networks*	*Ordered sets*
Internal dynamics	*Updates, propagation*	*Inferential reasoning*
Interaction	*Belief selection*	*Data mapping*

4.2 Belief Selection in DBR

Belief selection is the process that, given a specific set of available information, determines (1) what data are to be believed, and (2) which degree of strength is to be assigned to each of them, depending on the informational properties of the corresponding datum. In DBR, belief selection is handled by a mathematical system, including a *condition* C, a *threshold* k, and a *function* F. Condition C and threshold k together express the minimal informational requirements for a datum to be selected as belief, while the function F assigns a value of strength to the accepted beliefs (Paglieri, 2004; Paglieri and Castelfranchi, 2004). Let $\boldsymbol{B}$ represents the set of the agent's beliefs, and $B^s\varphi$ represents the belief φ with strength s. Hence the general form of the selection process is:

$C(c^\varphi, i^\varphi, l^\varphi) \leq k \rightarrow B^s\varphi \notin \boldsymbol{B}$

$C(c^\varphi, i^\varphi, l^\varphi) > k \rightarrow B^s\varphi \in \boldsymbol{B}$ with $s^\varphi = F(c^\varphi, i^\varphi, r^\varphi)$

The setting of C, F and k is an individual parameter, which might vary in different agents (cf. 4.5). Examples of individual variation in belief selection are the following:

$C: c^\varphi > k$	$k: 0.5$	$F: c^\varphi$
$C: c^\varphi > k$	$k: 0.6$	$F: (c^\varphi + i^\varphi + l^\varphi) / 3$
$C: c^\varphi > k \cdot (1 - l^\varphi)$	$k: 0.8$	$F: c^\varphi \cdot (i^\varphi + l^\varphi)$

All these parametrical settings assign to data credibility the main role in determining belief selection, but they do so in widely different ways. The first parametrical setting expresses a thoroughly realistic attitude toward belief selection, regardless of any considerations about importance or likeability. At the same time, the minimal threshold is set at a quite tolerant level of credibility (0.5). The threshold is slightly higher in the second parametrical setting, and the condition is identical: on the whole, this reflects a more cautious

acceptance of reliable data. But once a datum is indeed accepted as belief, its strength is now calculated taking into account also importance and likeability, in contrast to the previous setting. The same happens in the third parametrical setting, although along different lines. Here the threshold is extremely high (0.8), but the condition is influenced by likeability as well: assuming that likeability ranges in the interval **[0, 1]**, here the minimal threshold over credibility is conversely proportional to the likeability of the datum (e.g. it is 0.08 for a datum with likeability 0.9 vs. 0.72 for a datum with likeability 0.1). That expresses a systematic bias toward the acceptance of likeable (i.e. pleasant) data, in spite of their credibility. In other words, these parametrical settings define three agents with different attitudes, with respect to belief selection: a *tolerant full realist* (the first), a *prudent open-minded realist* (the second), and a *wishful thinking agent* (the third).

4.3 Data Structure and Information Update

In DBR, data structures are conceived as networks of nodes (data), linked together by characteristic relations. For the purposes of the present discussion, it will suffice to define three different types of data relations: support, contrast, and union:

I. *Support*: □ supports □ (□□ □) iff $c_{\square}$ $c_{\square}$, the credibility of datum □ is directly proportional to the credibility of datum □.
II. *Contrast*: □ contrasts □ (□□ □) iff $c_{\square}$ $1/c_{\square}$, the credibility of datum □ is conversely proportional to the credibility of datum □.
III. *Union*: □ and □ are united (□ &□) iff $c_{\square}$ and $c_{\square}$ jointly (not separately) determine the credibility of another datum □

Given a data structure, belief change is usually triggered by *information update* either on a fact or on a source: the agent perceives or infers a new piece of information, rearranges her data structure accordingly, and possibly changes her belief set, depending on the belief selection process. In DBR, information update specifies the way in which a new input is integrated in the agent's data structure (Fig. 3), emphasizing that such input (either external or internal) generates not only a new datum concerning its *content*, but also data concerning *source attribution* and *source reliability*, and the *structural relations* among them (Castelfranchi, 1997; Paglieri, 2004) – allowing, among other things, to implement in DBR sophisticated model of *trust assessment* (Fullam, 2003; Falcone and Castelfranchi, 2004).

The gist here is that the agent memory does not only store information, but it also keeps track of the way in which such information was acquired – and it is exactly because of these *'structural traces'* that the agent is usually capable to assess both her beliefs and the reasons behind them.

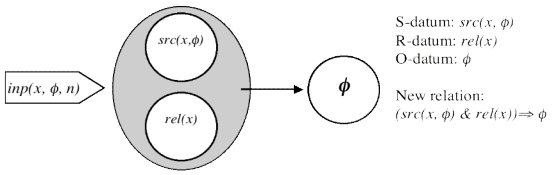


Figure 3: Information update in DBR

4.4 Assessment of Data Properties

While the assessment of credibility and importance is detailed in Paglieri (2004), here the focus is put on *relevance* and *likeability*, i.e. those informational properties which require a mapping between data and goals. In-depth technical discussion on the computational treatment of relevance and likeability is beyond the aim of this paper, so only the functional differences in DBR between these two kinds of mapping are discussed.

To start with, while the assessment of relevance involves every data *supporting* the pursued goal (e.g. data on means-end relations, know-how, urgency, feasibility, external conditions, etc.; Castelfranchi, 1995; 1998), the assessment of likeability requires only a mapping with the *contents* of the pursued goal, which by default puts a pressure (more or less influential, depending on the agent frame of mind; cf. 4.5 and 5) toward believing that the desired state of the world is indeed realized. This requires different mappings (i.e. functions) to assess these properties: in the case of relevance, the function will go from goals to sets of data[3], while likeability merely implies a comparison one-to-one between goals and data. By way of example, let us take an agent who pursues the goal of "being loved by his fiancé": this goal puts a pressure on believing that indeed his fiancé is in love with him (i.e. it makes the datum "being loved by his fiancé" more likeable to him), and in addition it defines a set of data which are relevant to such purpose, such as information on her tastes, schedule, future intentions concerning marriage, expectations about children, family relationships, and so on.

Once the mapping functions for a given agent are specified, the value of relevance of a certain datum at a given time is calculated by (1) checking of many currently pursued goals require such datum,

[3] Apart from the technical nuisance of effectively modelling such mapping, the true problem is to define a proper rationale for it: according to which criteria is a datum considered to be relevant for a certain goal? In future works, this issue will be tackled in DBR by trying to provide an operational counterpart of Castelfranchi's theory on the role of information in goal processing (Castelfranchi, 1997).

and (2) measuring the value of those goals for the agent – hence relevance is indeed proportional to the number and value of the goals for whom the datum is useful (cf. 4.1). The assessment of likeability is even more straightforward: if the content of the datum matches with the content of one of the agent goals, the likeability of the datum will be proportional to the value of the corresponding goal; otherwise, the datum likeability is set to zero.

Clearly enough, different agents might assess relevance and likeability in slightly different ways: e.g., some agents might be more thorough than others in screening useful data for a given goal (hence picking up a larger set of data as relevant for that purpose); similarly, given the same goal with the same value, two agents might differ in their evaluation of the bearing of such goal over belief formation (i.e. the corresponding datum might ends up with different value of likeability in the data structure of the two agents). As it is customary in DBR, these individual variations are handled by specifying agents with different parameters for the assessment of relevance and likeability (cf. 4.5).

4.5 Individual Variation in DBR

The DBR framework summarized here is based on a conceptual distinction between *principles* and *parameters*. Principles are *general* and *theoretical* in nature, defining the common features which characterize doxastic processing in every agent. Parameters, instead, are *individual* and *operational*, specifying in which fashion and measure each agent applies belief dynamics. The cognitive and social framework of the model is captured by its principles, while individual variation is represented through parametrical setting (Paglieri and Castelfranchi, 2004). Some examples were already discussed in 4.2, to show how parametrical variation can be effectively exploited to model *different attitudes in belief dynamics*. In particular, the application of individual parameters to motivational effects have been hinted at in 4.4, and it will be further developed in 5.

5 Relevance and Likeability: Motivations in DBR

Motivations are represented in this framework as pursued goals (Miceli and Castelfranchi, 1998; 2000): hence, relevance expresses the effects of *goal-pursuing* over belief selection, while likeability biases belief selection toward *goal-satisfaction*, i.e. conformity between how we consider the world to be, and how we would like it to be. More noticeably, in DBR relevance and likeability affect belief formation *at different stages and in different ways.* Relevance determines the sub-set of data taken in consideration by the agent at a given time (*what she sees*, so to speak), while likeability plays a direct role in the selection process which is being performed over these candidate data, hence influencing (to some extent) *what she believes*. With reference to Figure 2, relevance intervenes at the stage of *focusing*, while likeability might play a role in *belief selection*.

Another significant difference between these two motivational factors is that relevance exerts a *systematic* and *constant* influence over belief formation, while likeability affects belief selection only *occasionally* and *under specific conditions*.

One of the obvious trait of cognitive processing in general is to be resource-bounded, and the same applies to belief dynamics as well (Cherniak, 1986; Wassermann, 2000). Among other things, an agent does not select her current beliefs from the totality of the data stored in her memory, but rather from a limited sub-set of it: in DBR, such sub-set is focused on the basis of data relevance – hence the pervasive nature of relevance-based effects in belief dynamics.

In contrast, likeability might or might not affect belief selection, since it expresses a specific attitude toward motivation that not all agents need to share – nor the same agent is likely to be always equally biased toward confirmation of her expectations. Take for instance the selection policies detailed in 4.2: the first (full realism) is not influenced by likeability at all, while the second (open-minded realism) is affected to a very limited extent; only the third attitude (wishful thinking) is strongly biased by likeability effects – that is, influence of likeability depends on the agent parameters for belief selection.

Whatever their functional differences, both relevance-based effects (contextual influences over cognitive processing, tunnel vision, framing effects) and likeability-based effects (self-verification, avoidance of cognitive dissonance, pathological denial) are strongly supported by empirical evidence in several lines of research (Festinger, 1957; Kruglanski, 1980; Kahneman et al., 1982; Kunda, 1990; Swann, 1990; Oatley and Jenkins, 1996; Frijda et al., 2000; Swann et al., 2002). DBR tries to reproduce similar dynamics in a formal framework, with the long-term aim of implementation in artificial cognitive systems (Paglieri, 2004).

This yields a crucial question: are such influences *adaptive*, i.e. do they usually serve well the agent's practical purposes? And how is it so?

Answering such problems leads to further stress the *differences*, rather than the similarities, between relevance and likeability in affecting belief selection (Forgas, 1995; 2000). In fact, while relevance is clearly a prominent adaptive feature of information processing, serving to *avoid unnecessary*

computational costs for the assessment of useless information (Cherniak, 1986; Wassermann, 2000), the adaptive role of likeability is less obvious, and it rather relates to the agent's *defence mechanisms* on the one hand (Miceli and Castelfranchi, 1998), and to specific *heuristics* on the other (Todd and Gigerenzer, 2000). In the first case, likeability effects might help to preserve the affective balance of the agent, acting as homeostatic device to maintain a stable level of self-confidence in the face of an excess of unfavourable evidence: here a delicate trade-off is drawn between accuracy of beliefs and moderate degrees of wishful thinking, to avoid loss of motivation without plunging into utter self-delusion. As for the heuristic value of likeability, under certain conditions an overly optimistic attitude is indeed the most effective way of achieving a solution to the agent's need: for instance, in extreme situations optimism and stubbornness might help the agent to spot an opportunity as soon as it arises (e.g., water in the desert) – besides, accuracy of beliefs does not really matter here for all those agents that do not survive. Relevance effects show analogous heuristic value, although under quite different conditions: e.g., a maniacal focus over a very narrow set of data (such as in tunnel vision) is indeed desirable when the agent is fleeing from some danger or working under severe time constraints, to avoid being distracted.

Moreover, even when relevance and likeability fail to serve efficiently the agent purposes and affect negatively her performance, their characteristic biases are still remarkably different: *relevance-based biases* (e.g., tunnel vision, obsessive commitment, repeated failure in detecting an obvious solution) usually testify of some malfunctioning in the agent's goal-processing, rather than in belief formation dynamics – the agent is usually being obsessed by some specific motivation, so that she fails to shift her focus toward more profitable lines of reasoning. On the contrary, *likeability-based biases* (e.g., wishful thinking, pathological denial, self-delusion) generates from specific shortcomings in belief processing, either in the assessment of data properties (likeability of a datum is exaggerated, in comparison to the actual value of the corresponding goal), or in the belief selection process (too much weight is given to likeability, without paying enough attention to other crucial factors such as pragmatic credibility and epistemic importance).

Finally, it is quite interesting to discuss some experimental findings on motivated reasoning within the framework of DBR. In particular, different outcomes were observed between *accuracy-driven reasoning* and *goal-directed reasoning* (Kruglanski, 1980; Kruglanski and Ajzen, 1983; Kunda, 1990): in the first case, subjects were explicitly oriented toward a careful and dispassionate evaluation of the issue at hand for the sake of accuracy *per se*, while in the second case they were provided with a personal motive or interest to arrive at a particular, directional conclusion. Accuracy-driven reasoning was characterized by an increase in the quantity of belief processing (in terms of more alternatives being considered in a lengthier fashion; Kruglanski, 1980; Kruglanski and Ajzen, 1983), but also by a different quality of such processing, with a tendency to use more sophisticated and deliberate inferential procedures (Kunda, 1990) – a tendency that DBR would capture as a *parametrical switch* (cf. 4.5), with a fine-tuning of the heuristics applied in belief selection. In contrast, goal-directed reasoning showed exactly the kind of biases discussed so far, with belief processing constrained by the data taken into account by the agent (relevance-based biases on focusing; Kruglanski, 1980; Kruglanski and Ajzen, 1983) and by a pressure toward confirmation and self-verification (likeability-based biases on belief formation; Festinger, 1957; Swann, 1990).

However, an additional feature emerged quite clearly in goal-directed reasoning, as remarked by Kunda: «The biasing role of goals is […] constrained by one's ability to construct a justification for the desired conclusion: people will come to believe what they want to believe only to the extent that reason permits» (1990: 483). This seems to suggest that direct influence of likeability on belief formation is usually only *provisional* and *temporary*, since it needs to be backed up by factual supports (i.e. credibility and importance, in DBR terms) as soon as possible. Moreover, human agents seem to be inclined to *actively search for such supports and justification*, performing specific epistemic actions to this purpose. This opens interesting scenarios for further research on the role of motivation in belief dynamics (cf. 7).

6 Passionate Believers: Emotions in DBR

Although so far the modelling of DBR was mainly aimed to integrate belief dynamics with *goals*, hence motivations, some connections with the related field of emotions can tentatively be highlighted – merely as suggestions for future work (cf. 7). At present, the effects of emotions over DBR belief processing seem to show mainly as:

I. *Motivation inducement*: this is an indirect and local effect, due to the fact that emotions are well-known motivational engines (Ortony et al., 1988; Frijda et al., 2000), i.e. they often generate and/or trigger specific motivations in the agent's

mind. In turn, such motivations affect the agent's beliefs (cf. 5), and it is possible to map the significant effects of the triggering emotion on the doxastic dynamics of the agent: for instance, a sudden feeling of uneasiness and fear, by inducing the belief that there is some unknown danger ahead, strongly drives the agent's attention toward means of escape, forcing her to neglect other available information.

II. *Parametrical readjustment*: this is a direct and pervasive effect, in which a certain emotion affects and modifies the whole processing of the agent. In DBR, this is effectively captured by a readjustment of one or more of the agent's parameters: for instance, in some subjects fear can induce a lowering of the selection threshold (i.e., they are ready to take action on less reliable bases), but also a greater role for credibility in both condition and function (i.e., they become much more concerned with factual assessment of the best course of action, rather than epistemic importance or likeability; cf. 4.2).

7 Conclusions and Future Work

This work presented a preliminary attempt of integrating motivation in the formal framework of DBR, to spell out the influence of goals over procedural belief formation, and the different role played by relevance and likeability in such influence. However, this work is still in progress, and several aspects remain to be explored in detail.

In particular, future researches will aim to refine the DBR model and to move toward implementation in agent-based cognitive and social simulation (Dragoni and Giorgini, 2003; Paglieri, 2004), to focus on motivational and emotional effects over belief formation in a broader perspective (Clore and Gasper, 2000; Forgas, 2000), and to investigate the interplay between beliefs, motivations and emotions in social interaction, e.g. in argumentation (Paglieri and Castelfranchi, 2004).

Acknowledgements

I wish to thank Cristiano Castelfranchi and Maria Miceli for insightful and enjoyable discussion on these topics. Special thanks are also due to the excellent anonymous reviewers of my extended abstract: their criticisms and suggestions greatly improved the final version. This work was developed at the ISTC-CNR in Roma, as part of the PAR 2003 project *Prospective Minds: Toward a theory of anticipatory mental representations*, financed by the University of Siena.

References

Alchourrón, C., P. Gärdenfors and D. Makinson: 1985, 'On the logic of theory change: Partial meet Contraction and revision functions', *Journal of Symbolic Logic* **50**, 510-530.

Anderson, J. R.: 1996, 'ACT: A simple theory of complex cognition', *American Psychologist* **51**, 355-365.

Anderson, J. R., D. Bothell, M. D. Byrne, S. Douglass, C. Lebiere and Y. Qin: 2004, 'An integrated theory of the mind', *Psychological Review* **111**, 1036-1060.

Berger, J. O.: 1985, *Statistical decision theory and Bayesian analysis*, Springer-Verlag, New York.

Boutilier, C.: 1998, 'A unified model of qualitative belief change: a dynamical systems perspective', *Artificial Intelligence* **98**, 281-316.

Cañamero, L.: 2003, 'Designing emotions for activity selection in autonomous agents', in R. Trappl, P. Petta and S. Payr (eds.), *Emotions in Humans and Artifacts,* The MIT Press, Cambridge MA, 115-148.

Castelfranchi, C.: 1995, 'Guarantees for autonomy in cognitive agent architecture', in M. Wooldridge and N. Jennings (eds.), *Intelligent Agents*, Springer-Verlag, Berlin, 56-70.

Castelfranchi, C.: 1996, 'Reasons: Belief support and goal dynamics', *Mathware & Soft Computing* **3**, 233-247.

Castelfranchi, C.: 1997, 'Representation and integration of multiple knowledge sources', in Cantoni, Di Gesù, Setti and Tegolo (eds.), *Human & machine perception*, Plenum Press.

Castelfranchi, C.: 1998, 'Modelling social action for AI agents', *Artificial Intelligence* **103**, 157-182.

Cherniak, C.: 1986, *Minimal rationality*, The MIT Press, Cambridge MA.

Clore, G. and K. Gasper: 2000, 'Feeling is believing: Some affective influences on belief', in N. Frijda, A. Manstead and S. Bem (eds.), *Emotions and beliefs: How feelings influence thoughts*, CUP, Cambridge MA, 10-44.

Doyle, J.: 1992, 'Reason maintenance and belief revision: Foundations vs. coherence theories', in P. Gärdenfors (ed.), *Belief revision*, CUP, Cambridge MA, 29-51.

Dragoni, A. F. and P. Giorgini: 2003, 'Distributed belief revision', *Autonomous Agents and Multi-Agent Systems* **6**, 115-143.

Evans, D., A. Heuvelink and D. Nettle: 2003, 'The evolution of optimism: A multi-agent based model of adaptive bias in human judgement', in *AISB'03 Symposium on Scientific Methods for the Analysis of Agent-Environment Interaction*, University of Wales, 20-25.

Fagin, R. and J. Halpern, 1994: 'Reasoning about knowledge and probability', *Journal of the ACM* **41**, 340-367.

Falcone, R. and C. Castelfranchi: 2004, 'Trust dynamics: How trust is influenced by direct experiences and by trust itself', in N. Jennings, C. Sierra, L. Sonenberg, M. Tambe (eds.), *Proceedings of AAMAS'04*, ACM, 740-747.

Festinger, L.: 1957, *A theory of cognitive dissonance*, Stanford University Press, Stanford CA.

Forgas, J.: 1995, 'Mood and judgement: The affect infusion model (AIM), *Psychological Bulletin* **117**, 39-66.

Forgas, J. (ed.): 2000, *Feeling and thinking: The role of affect in social cognition*, CUP, Cambridge MA.

Friedman, N. and J. Halpern: 1999, 'Belief revision: A critique', *Journal of Logic, Language and Information* **8**, 401-420.

Frijda, N., A. Manstead and S. Bem (eds.): 2000, *Emotions and beliefs: How feelings influence thoughts*, CUP, Cambridge MA.

Fullam, K.: 2003, *An expressive belief revision framework based on information valuation*, MS thesis, University of Texas at Austin.

Gärdenfors, P.: 1988, *Knowledge in flux: Modelling the dynamics of epistemic states*, The MIT Press, Cambridge MA.

Huns, M. and D. Bridgeland, 1991: 'Multi-agent Truth Maintenance', *IEEE Transactions on Systems, Man and Cybernetics* **21**, 1437-1445.

Kahneman, D., P. Slovic and A. Tversky: 1982, *Judgment under uncertainty: Heuristics and biases*, CUP, Cambridge.

Kruglanski, A.: 1980, 'Lay epistemology process and contents', *Psychological Review* **87**, 70-87.

Kruglanski, A. and I. Ajzen: 1983, 'Bias and error in human judgement', *European Journal of Social psychology* **13**, 1-44.

Kunda, Z.: 1990, 'The case for motivated reasoning', *Psychological Bulletin* **108**, 480-498.

Miceli, M. and C. Castelfranchi: 1998, 'Denial and its reasoning', *British Journal of Medical Psychology* **71**, 139-152.

Miceli, M. and C. Castelfranchi: 2000, 'Nature and mechanisms of loss of motivation', *Review of General Psychology* **4**, 238-263.

Oatley, K. and J. Jenkins: 1996, *Understanding emotions*, Blackwell Publishing, Oxford.

Ortony, A., G. Clore and A. Collins: 1988, *The cognitive structure of emotions*, CUP, New York.

Paglieri, F.: 2004, 'Data-oriented Belief Revision: Toward a unified theory of epistemic processing', in E. Onaindia and S. Staab (eds.), *Proceedings of STAIRS 2004*, IOS Press, Amsterdam, 179-190.

Paglieri, F. and C. Castelfranchi: 2004, 'Revising beliefs through arguments', in I. Rahwan, P. Moraïtis and C. Reed (eds.): *Argumentation in Multi-Agent Systems*, Springer, Berlin.

Picard, R.: 1997, *Affective Computing*, The MIT Press, Cambridge MA.

Pollock, J. and A. Gillies: 2000, 'Belief revision and epistemology', *Synthese* **122**, 69-92.

Rott, H.: 2001, *Change, choice and inference: A study of belief revision and nonmonotonic reasoning*, Oxford University Press, Oxford.

Segerberg, K.: 1999, 'Two traditions in the logic of belief: Bringing them together', in H. J. Ohlbach and U. Reyle (eds.), *Logic, language and reasoning*, Kluwer, Dordrecht, 135-147.

Swann, W.: 1990, 'To be adored or to be known? The interplay of self-enhancement and self-verification', in R. M. Sorrentino and E. T. Higgins (eds.), *Foundations of social behavior*, Guilford, New York, vol. 2, 408-448.

Swann, W., P. Rentfrow and J. Guinn: 2002, 'Self-verification: The search for coherence, in M. Leary and J. Tagney (eds.), *Handbook of self and identity*, Guilford, New York, 367-383.

Todd, P. and G. Gigerenzer: 2000, 'Simple heuristics that make us smart', *Brain and Behavioural Sciences* **23**, 727-741.

Wassermann, R.: 2000, *Resource-bounded belief revision*, ILLC DS-2000-01, Amsterdam.

Cost minimisation and Reward maximisation. A neuromodulating minimal disturbance system using anti-hebbian spike timing-dependent plasticity.

Karla Parussel
*Department of Computer Science and Maths,
University of Stirling,
Stirling, FK9 4LA, Scotland
kmp@cs.stir.ac.uk

Leslie S. Smith
†Department of Computer Science and Maths,
University of Stirling,
Stirling, FK9 4LA, Scotland
lss@cs.stir.ac.uk

Abstract

In the brain, different levels of neuro-active substances modulate the behaviour of neurons that have receptors for them, such as sensitivity-to-input, Koch (1999). An artificial neural network is described that learns which actions have the immediate effect of minimising cost and maximising reward for an agent. Two versions of the network are compared, one that uses neuromodulation and one that does not. It is shown that although neuromodulation can decrease performance it agitates the agent and stops it from over-fitting the environment.

1 Introduction

Fellous (1999) proposes that emotion can be seen as continuous patterns of neuromodulation of certain brain structures. It is argued that theories considering emotions to emanate from certain brain structures and from non-localised diffuse chemical processes should be integrated. Three brain structures are considered in this way; the hypothalamus, amygdala and prefrontal cortex.

Fellous (2004) further suggests that the focus of study should be on the *function* of emotions rather than on what they are. Seen in this way, animals can be seen functionally as having emotions, whether or not we empathise with them. Given this, robots can functionally have emotions as well. One function of emotions mentioned that has a robotic counterpart is to achieve a multi-level communication of simplified but high impact information.

One way of studying the functionality of emotions, is to identify the extra functionality provided by neuromodulation compared to a non-modulating solution. Modulation is used here to signal agent needs in a neural network that is used for the purpose of action selection. The structures of both solutions are inherently the same but the modulating version has the added interaction between neuromodulation and neural network.

Although neuromodulators and hormones have been emulated for the purpose of action selection before, Avila-Garcia and Canamero (2004) Shen and Chuong (2002), they have not been applied to a purely neural network solution and have not been compared to non-modulating versions. Husbands (1998) evolves controllers inspired by the modulatory effects of diffusing NO. This speeds up evolutionary production of successful controllers.

The difficulty is that what can be done with a modulating network, can also be done with a non-modulating network if it has been evolved for that purpose. Therefore the comparison needs to be made in an environment that the agent has not been evolved for.

2 Application of the model

An adaptive agent needs a reason to adapt in order to do so. A common reason is to maximise and retain resources. In this context a resource is a single continuous scalar value that correlates with a characteristic of the state of the agent or environment. A resource can correlate with a single quantifiable level such as a battery charge level for a physical robot, or be an estimation of a virtual non-measurable level such as utility or safety. An adaptive agent is faced with two tasks when maximising these resources, that of learning to perform actions which result in an increase in a resource level, and that of learning not to perform actions which result in a decrease of resource.

Here, the Artificial Life animat concept is abstracted to provide the simplest possible context for testing the effect of neuromodulation applied to an artificial neural network. The agents have no external senses to adapt to and can only sense their internal state. The choice of output directly and immediately alters the internal state of the agent, which therefore alters the strength of the input signal to the network.

The agent has a body that requires two resources, energy and water. It keeps track of the largest increase and decrease of each. The current change in resource level is then scaled to these maxima to be within the range [0,1] before being passed to the network.

The agent is given a set of actions that increase or decrease by one or two resource points[1], or are neutral to, either the energy or water level in the body. There is one action for each permutation making 10 in total.

3 Implementation

3.1 Topology

The network consists of three layers of adaptive leaky integrate and fire neurons learning via spike timing-dependent plasticity, G. Q. Bi (2002). The model learns which outputs should be most frequently and strongly fired to minimise the level of input signal. There is one output neuron per action. The action has an effect on the internal state of the agent, which determines the strength of the input signal to the network. For the modulating network, the input layer neurons increase modulator strengths when fired, while the middle layer neurons have receptors for those modulators.

There are situations in which an effective behaviour for an agent may decrease a need but not satisfy it. For example, if it is in an environment which is temporarily bereft of resources then waiting and conserving its current levels may be the optimal behaviour. Alternatively there may be situations in which an agent needs to store more resources than it normally does. In this case the need for the resource will be signalled despite that need being signalled as satisfied. An example would be an agent expecting to find itself in an environment bereft of resources.

For each resource the input layer has two neurons that output to the middle layer. One neuron signals the need for the resource and the other neuron signals the satisfaction of that need. If a previous action performed by the agent results in a decrease in hunger or an increase in resource satiation, then the corresponding input signal is momentarily decreased.

The model uses a feed forward network that can be iterated over a number of times within a single turn, after which the winning output neuron is chosen. Which neuron wins is determined by summing up the total charge of each neuron over all the iterations and choosing the neuron with the greatest sum. This stops a neuron with strong inputs from losing because it just has spiked and thus has low activity or is in a refractory period.

3.2 Modulators

Two variants of the network were created; modulating and non-modulating. They were the same except that the modulating network had in addition two modulators, one used to signify hunger and the other thirst.

As used here, a modulator is a global signal that can influence the behaviour of a neuron if that neuron has receptors for it. The signal decays over time, specified by the re-uptake rate, and can be increased by firing neurons that have secretors for it.

Neurons within the middle layer are given a random number of receptors. These can be modulated by neurons in the input layer that have secretors. These neurons were given a random number of secretors. The receptors modulate either the neuron's sensitivity to input or probability of firing. The extent of this is determined by the level of the associated modulator and whether the receptor is inhibitory or excitatory. The secretors increase an associated modulator. The modulation rate of the receptors and the increment rate of the secretors is determined by evolution.

4 Parameter Optimisation

The network has a number of parameters which must be set correctly for it to adapt successfully. These are parameters that have no obvious value, such as the number of neurons in the middle layer, secretion, modulation rates etc. Automated parameter optimisation was performed for the modulating and non-modulating agents. Afterwards the parameter sets were hard-coded and tests were performed upon a population of agents using them.

The fitness of an agent was determined by

$$Energy + Water + Age - |Energy - Water|$$

The difference between the energy and water resource was subtracted from the fitness as both resources were essential for the agent to stay alive.

[1] Points are used as it is an arbitrary level that has no correlation with any real physical quantity.

5 Results

After optimisation, a modulating and a non-modulating agent were picked for further testing. The fitness of the genotypes were equivalent and both were typical of the solutions that were evolved. Because there was a stochastic mapping from genotype to phenotype and to provide multiple evaluations, the agents were hard-coded so that they could be tested as a population.

Parameter optimisation converged upon a fully hebbian network for the non-modulating network and a hybrid anti-hebbian / hebbian network for the modulating network.

5.1 Initial tests

When viewed over the course of the agent lifetime it can be seen that a typical agent learns which actions provide minimal disturbance to its inputs. It initially chooses a neutral action before settling on the most rewarding water action. The agent then alternates between this and the most rewarding energy action; see figure 1. Figure 2 shows the initial learning process before one output neuron wins over all the others.

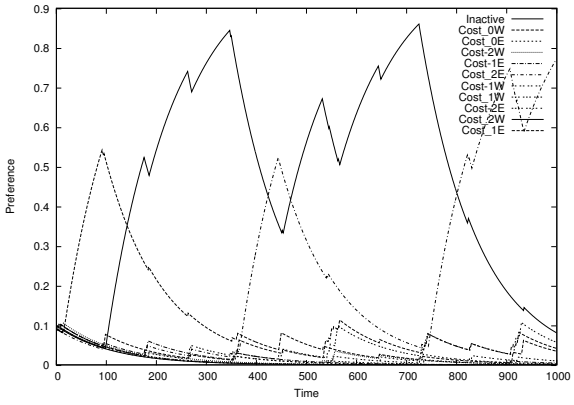


Figure 1: Actions chosen over lifetime of a single modulating agent.

The performance of the non-modulating and the modulating agents were similar although on average the non-modulating network would reach higher levels of fitness and would be optimised by the parameter search more quickly.

5.2 Extended tests

During parameter optimisation, each genotype was tested for 1,000 turns before being evaluated. The evaluation was cut short if the agent died prematurely because a resource had decreased to nothing.

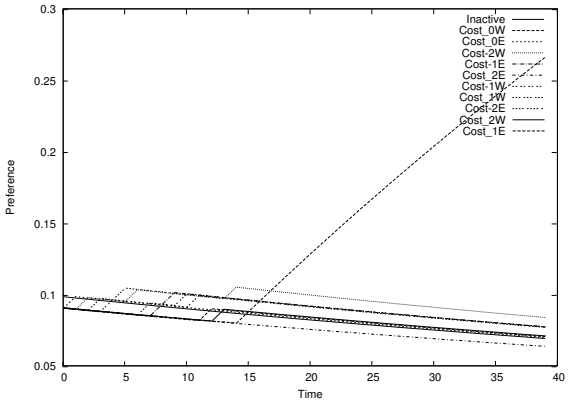


Figure 2: The first 40 cycles of the run in figure 1 showing the initial learning process.

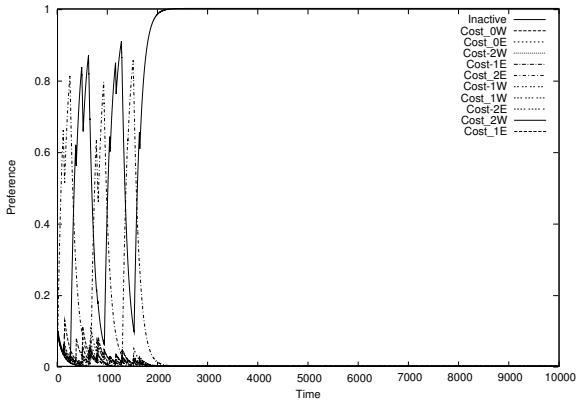


Figure 3: Non-Modulating agent run over an extended period of time (10,000 turns).

After parameter optimisation, when testing a population of hard-coded non-modulating agents for longer than 1,000 turns, the activity in the network ceased over time. The charge of the output neurons would slowly decay over time with the winning action remaining the same each time; see figure 3.

The limited use of artificial evolution for parameter optimisation had settled upon a brittle strategy which depended on how long each agent was evaluated for.

A population of hard-coded modulating agents were then tested for the same extended period of time. They were shown to continue transitioning between the same two winning output neurons that caused a maximum increase in energy and water, with other neurons very occasionally being chosen; see figure 4. Modulation had prevented the artificial evolution

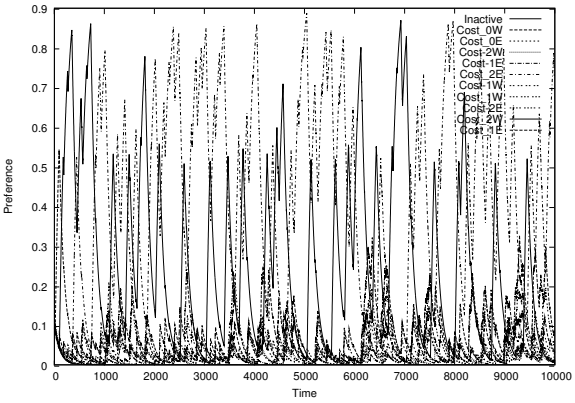


Figure 4: Modulating agent run over same extended period of time.

used for the parameter optimisation from over-fitting the test environment.

6 Discussion

It was discovered that the network performed most effectively when the actions it chose could minimise input activity. Wörgötter and Porr (2004) provide an overview of the field of temporal sequence learning. They discuss how the learning paradigm of disturbance minimisation, as opposed to reward maximisation, removes the problem of credit structuring and assignment. The two paradigms are not equivalent. Whereas maximal return is associated with a few points on a decision surface, minimal disturbance uses all of the points. Every input into the system drives the learning and when there are no inputs then the system is in a desirable, stable state.

Modulation agitates the network, stopping it from settling into a stable state for too long or letting activity decline to a point whereby the network stops alternating between actions. When tested using an extended run, the modulating network, unlike the non-modulating version, continues to alternate between the actions causing the least input disturbance throughout its lifetime. Figure 4 shows that other actions always have a chance of being selected.

When comparing the modulating and non-modulating agents in environments that they were not evolved for, in this case evaluated for an extended length of time, then it is shown that modulation makes the agent more robust. This robustness carries with it a performance cost.

This suggests that one functional use of emotions is to provide agitation to the agent in order to not let it settle into a stable state. Even though the environment may allow for it or make this the optimal behaviour. An explanation for this could be that natural agents have not evolved for such environments because they rarely exist and cannot be relied upon to last.

7 Acknowledgements

Karla Parussel acknowledges the financial support of the EPSRC and the University of Stirling. The authors wish to thank the anomynous referees for their useful comments.

References

O. Avila-Garcia and L. Canamero. Using hormonal feedback to modulate action selection in a competitive scenario. In *From Animals to Animats 8: Proceedings of the eighth international conference on the simulation of adaptive behavior*, pages 243–252. MIT Press, 2004. ISBN 0262693410.

Jean-Marc Fellous. The neuromodulatory basis of emotion. *The neuroscientist*, 5(5):283–294, 1999.

Jean-Marc Fellous. From human emotions to robot emotions. In *Architectures for Modeling Emotions: Cross-Disciplinary Foundations. Papers from the 2004 AAAI Spring Symposium*, pages 37–47. AAAI Press, 2004.

H. X. Wang G. Q. Bi. Temporal asymmetry in spike timing-dependent synaptic plasticity. *Physiol Behav.*, 77(4-5):551–555, 2002.

P. Husbands. Evolving robot behaviours with diffusing gas networks. In P. Husbands and J.-A. Meyer, editors, *Evolutionary Robotics: First European Workshop, EvoRobot98*, pages 71–86. Springer-Verlag, April 1998.

Christof Koch. *Biophysics of Computation*. Oxford University Press., 1999. ISBN 0-19-510491-9.

Wei-Min Shen and Cheng-Ming Chuong. The digital hormone model for self-organization. In *From animals to animats 7: Proceedings of the seventh international conference on simulation of adaptive behavior*, pages 242–243. MIT Press, 2002. ISBN 0-262-58217-1.

F. Wörgötter and B. Porr. Temporal sequence learning, prediction and control - a review of different models and their relation to biological mechanisms, 2004.

Motivating Dramatic Interactions

Stefan Rank*
*Austrian Research Institute
for Artificial Intelligence
Freyung 6/6, A-1010 Vienna, Austria
stefan.rank@oefai.at

Paolo Petta*†
†Dept. of Medical Cybernetics and Artificial Intelligence
of the Centre for Brain Research
at the Medical University of Vienna
Freyung 6/2, A-1010 Vienna, Austria
paolo.petta@meduniwien.ac.at

Abstract

Simulated dramatic story-worlds need to be populated with situated software agents that act in a dramatically believable way. In order to provide flexible roleplayers, agent architectures should limit the required external macro-level control. We present work on an architecture that exploits social embedding and concepts from appraisal theories of emotion to achieve the enactment of simple cliché plots. The interplay of motivational constructs and the subjective evaluative interpretation of changes in an agent's environment provide for the causal and emotional connections that can lead to the unfolding of a story.

1 Introduction

This work is part of the ActAffAct project (Acting Affectively affecting Acting (Rank, 2004)) that researches a bottom-up approach to imitating emotional characters that interact in a story-world. The goal is to achieve the unfolding of a plot-like structure while limiting the use of external macro-level control—as exerted by, e.g., a director. The ideal level of external control would be none at all, resulting in the emergence of plot from the characters' interaction, effectively turning the agents into reusable roleplayers. The question that arises is what are the **motivating elements** in the control architecture of synthetic characters that can provide for a dramatically appropriate sequence of actions. Our approach views emotions—as described in appraisal theories (Frijda, 1986; Scherer et al., 2001; Ortony, 2003)—as the links between actions that render a plot plausible.

Emotions are the essence of a story (Elliott et al., 1998) and play a central role in engaging drama. The conflicts between the characters in a play and the emotions involved in resolving them are the constituents of a dramatic structure, a plot. **Drama** can be described as the art that deals with a refined version of emotional interaction between individuals (Vogler, 1996; Egri, 1946). These ideas provide a starting point and can serve as success criteria for the creation of **dramatic story-worlds**, i.e., simulations that are inhabited by software agents for the purpose of enacting dramatically interesting plots. These worlds present themselves to the single agent as inherently social domains, as social interaction is often crucial for solving problems. In the ideal case the author of such a story-world would be able to shift from today's specification of exact sequences of actions to the authoring of possible actions, regularities in the environment, and the setting up of an initial constellation of characters, including their general traits. This would not necessarily be an easier process of creation but it could lead to a more flexible, and possibly user-driven, experience of dramatic structures.

Using an **appraisal-based architecture** that considers the social and physical lifeworld of an agent is seen as key to construct emotionally and dramatically believable characters for interactive drama. This paper highlights aspects of our extensions to a BDI architecture (belief, desire, intention (Bratman et al., 1988; Huber, 1999)) that are pertinent to motivating dramatic actions in a minimal version of a cliché storyline. Situated appraisal of all percepts, a three-phase model of behaviours, and varied coping behaviours have been our first steps towards character-based narrative.

2 Enacting the Social Lifeworld You are Embedded In

Contrary to its physical surroundings, the intangible **social lifeworld** an agent is embedded in has to be continually enacted and negotiated. We use the term social lifeworld in the tradition of Agre and Horswill's analytic endeavour (Agre and Horswill, 1997)

to thicken the notion of environment. Our focus on sociality combines the goal of reducing the cognitive load for individual entities populating the environment with emphasising the relevance of coordinative functions (Clancey, 1999) mediating between an individual and the (potential) current and future opportunities and threats to satisfy an individual's **concerns** (Frijda, 1986).

The notion of concern is defined as subjective disposition to desire occurrence or nonoccurence of a given kind of situation. This definition, taken from (Frijda, 1986), is related to but distinct from goals and motives, as the latter terms induce connotations of activity control. Concerns range from very concrete considerations—i.e., relating to an agent's immediate tasks—to abstract ones—such as feeling competent—that can lie dormant until an emotionally pertinent event takes place. The process of appraisal is described as a fast and possibly only partial evaluation of subjective significance of changes in the environment according to specific criteria.

By operationalising these theoretical notions, we de-emphasise the role of high-level cognition ("thinking") in routine functioning (Bargh and Chartrand, 1999) and recognise the opportunity offered by available structures to constrain high-level function within tractable bounds. Agre and Horswill (1997) identify abstract locatedness and functionally significant relationships grounded in the physical environment. Analogously, social lifeworld analysis considers the potential for *inter*-action with respect to loci of control at the macro level (e.g., power and status (Kemper, 1993)) as well as indirect access to (second and higher level) resources (e.g., Aubé, 1998).

A situated agent's dependence on **regularities** thus extends beyond the physical world into socio-cultural constructs, whose maintenance can e.g. be modelled as an interplay of conventions (social norms) and evaluative processes (emotions) (Staller and Petta, 2001), with emotions sustaining social norms, and culturally defined social norms in turn shaping and regulating emotions (e.g., with feeling rules defining which emotions are suitable in which situation, and display rules providing repertoires of how to express them).

Interpretations of situations and developments in the social lifeworld are not a given: both within an individual and in the society, they are the outcome of negotiations and transactions, captured e.g. in sociological models (Kemper, 1993) or characterised in terms of personality traits. Purposeful functions such as threatening, sanctioning, and amending, therefore are intrinsic behavioural requirements, along with their affective grounding in the social lifeworld. Emotional processes (ranging from raw affect under rough and undifferentiated circumstances, over fleets of feelings in (yet) unclear scenarios, to fully articulated "emotions" as result of detailed perception) mediate the translation between the subjective worlds of states, concerns and preferences, the abstract enacted shared social lifeworld, and the status and offerings of the physical world.

The enactment of the social lifeworld is the sphere of activity that is dominant in the context of dramatic interactions. The mechanisms of emotional processes for interpreting and sustaining this lifeworld and their influence on motivation form a fertile ground for drama.

3 Appraisal-based Architecture

It is a big step from the qualitative appraisal theories to an actual implementation. Several theoretical efforts investigate agent architectures that incorporate ideas about emotions (Isla et al., 2001; Frankel, 2002; Sloman and Scheutz, 2002; Gratch and Marsella, 2004; Marinier and Laird, 2004). The architecture we implemented uses ideas of TABASCO (Petta, 2003), the implementation effort has been based on JAM, the Java Agent Model (Huber, 2001). As a BDI architecture, JAM provides a plan representation language, goal- and event-driven (i.e., proactive and reactive) behaviour, a hierarchical intention structure, and utility-based action selection.

Figure 1: Four characters in ActAffAct

For our simple story-world we built a simulation including a graphical representation of an environment inhabited by four agents, taking on the roles of **narrative archetypes**: a hero, an antagonist, a mentor, and a victim (Figure 1). To provoke dramatic conflict, the agents are initialised with conflicting top-level goals (as a first approximation of concerns) and the social lifeworld is filled with entities suitable for creating and resolving said conflicts. The top-level goals include "being loved by someone" and "being mean to lovers", examples of dramatic entities—besides the agents themselves—are a flower; a sword; and the key to a treasure. The JAM model was adapted for concurrent execution and asynchronous interaction with this world. Apart from these surface changes, the architecture needed to be extended to support appraisal of perceptions in relation to the current goals of the agent, as these are the motivating structures in JAM. Furthermore, our use of plans in JAM was specifically tailored to the needs of an appraisal-based agent. As described in the next section, we restricted the flexible hierarchy of plans in JAM's intention structure to defined levels and implemented a further type of plan suitable for a situated perception process. Percepts are represented as JAM facts; as part of the appraisal process, however, they are reinterpreted by the agent according to its situated context.

According to the revised OCC model (Ortony, 2003) that incorporates more of the elements as discussed in e.g. Frijda (1986), appraisal is based on goals, standards, and preferences of the individual. The latter two are missing in JAM as explicit entities, but can be represented as beliefs. A separate appraisal component was added in the sequential execution cycle of the agent to perform the constant evaluation of percepts. The next section discusses in more detail elements of this architecture and their pertinence to appraisal and to motivating believable behaviour in a character.

4 Motivating Elements

Among our changes to the BDI model of JAM are the following additions and restrictions:

- **Perceptions plans**:

  These restricted plans are executed for matching percepts when they are first perceived. They implement the situated reinterpretation of percepts, translating from an agent-neutral representation to one that takes the agent's current context into account, and can range from asserting that an object near the agent is reachable to interpreting the picking up of a flower by the agent next to me as the anticipation of the possibility of being offered a present, thereby forming an expectation.

  This interpretion in the current context can be seen as the first step of appraising the significance of an event. It already takes into account components necessary for the social aspect of appraisal, such as determining the agent reponsible for a specific change.

- **Plan levels**:

  The space of plans accessible to an agent is structured in a hierarchy, starting from *concerns* at the top level, longer-term *activities*, and *behaviours*, to simple *action packages* and plans dedicated to executing a single *act*. This restriction in the use of hierarchy in JAM was chosen as plans of a given level share characteristic patterns. Behaviours in particular have been specifically designed to allow a simple and tractable implementation of appraisal. They are categorised as either trying to achieve something, helping somebody else to achieve something, or hindering them from achieving it. This reduces a part of the task of cognitive appraisal—namely assessing the relevance of a percept to one's own goals—to simple pattern matching (although more complex forms of relevance assessment are possible and desirable). The same holds for assessing the conformance of an action to the standards of an agent, e.g., the social norms, as these are expressed in terms of behaviours as well.

- **Behaviour phases**:

  Furthermore the execution of behaviours has been split into three phases of which the first and the last one are hard to interrupt, in order to simulate commitment to one's intentions. (In contrast, the utility-driven reasoning of JAM might possibly drop a just-started behaviour as well as one that is near its successful completion). A timed pattern was used for behaviours that influences the execution depending on the level of completion. In current work on regulatory influences on plan execution this capability is considered as part of a meta-level plan.

- **Expression and coping plans**:

  If the numerical intensity of the appraisal of a perceived and interpreted fact exceeds a certain

threshold value it creates a goal to cope with this situation and another one to express the agent's state. Expressive actions that indicate an agent's emotional state are in turn perceived and interpreted by other agents and trigger appraisals. This signalling of the current emotional state of an agent thus serves the purpose of revealing the elicitation of an emotion directly to others that are watching (Reisenzein, 2001). Coping introduces new top-level goals that are the main source of variation in generated plots. Coping activities motivate action that, by way of the emotion process, is causally related to percepts and concerns of the agent. These plans use the information made available by the appraisal of an event to decide on a suitable course of action to tackle the subjective interpretation. Overall, this provides for the causal relations needed for a dramatic plot.

ActAffAct's simulated domain was tested with different setups of the four characters, one of which excluded the antagonist and thereby the main source of conflict. In the no-conflict case the resulting interactions of the characters, not surprisingly, cannot be described as dramatic. A qualitative evaluation of the scenario with the full cast, however, leads us to believe that minimal storylines can indeed be generated using our approach. A quantitative evaluation was not yet pursued as this would require a measure of "storyness" for the automatic comparison of generated sequences of action, a complex research problem on its own (Charles and Cavazza, 2004). Even so, our approach shows that rather simple emotional extensions of a BDI architecture can yield reasonable outcomes in the distinctively social domain of dramatic interaction.

5 Related Work

Several recent projects that include simulated worlds target the area of interactive narratives in a wide sense (Magerko et al., 2004; Mateas and Stern, 2002; Cavazza et al., 2002), others pursue pedagogical applications (Machado et al., 2001; Marsella et al., 2000). A common problem of both types is the **narrative paradox**, the need to balance the flexibility of such a world with the control about narrative flow. As stated above, ActAffAct is designed taking the rather extreme viewpoint that external control can be reduced substantially without abandoning the claim of dramatically appropriate interactions. The crucial point is, to our mind, the reliance on emotional processes to provide the causal structure of action sequences.

EMA (Gratch and Marsella, 2004) is a framework for modelling emotion that tries to be domain-independent by harnessing concepts from appraisal theories of emotion. In EMA, coping is defined as inverse operation of appraisal, i.e., the identification and influencing of the believed causes for what has been appraised as significant in the current context. The main focus of development currently lies on extending the range of coping strategies (e.g., "mental disengagement", "positive reinterpretation", "further assess coping potential", or "planning") as responses to emotionally significant events.

Haunt2 (Magerko et al., 2004) is an attempt to create a game in which AI characters are central to the game experience. It is realised as a "mod" for the Unreal game engine. The goal of the game is to escape a house by influencing other characters indirectly. The dramatic storyline in Haunt2 is predefined, represented as a kind of partially ordered plan used by an explicit AI director to send commands to the different characters while reacting to unexpected moves by the human player.

A similar approach is used in **Façade**[1], where the proclaimed goal is interactive drama in a realtime 3D world (Mateas and Stern, 2002). Its public release is (at the moment) announced for spring 2005. In Façade there is also a separate component, external to the story, that arranges story segments ("beats") into a coherent story. The characters themselves act autonomously but adhere to the constraints of the established current story context. The ActAffAct project, in contrast, tries to achieve a simpler but similar effect without external control.

6 Conclusion And Further Work

Although we cannot yet claim to have succeeded in creating a robust generator of narratives, we nevertheless think that the approach of using emotional concepts in the control architecture of dramatic characters holds great promises to enrich dramatic storyworlds. We currently plan on integrating explicit regulatory strategies into the control architecture of the agents. The main focus of the effort to implement emotion regulation is to strengthen the coherence of a single agent's actions over longer time periods. In the context of an effort carried out within the European Network of Excellence HUMAINE[2], a broader

[1] http://www.interactivestory.net/#facade

[2] http://emotion-research.net

survey work and steps towards a principled approach for the integration of affective processes, deliberation, and situated action in viable agent architectures are being undertaken. The long term goal is to clarify the systematic relation between the complexity of an environment including its social characteristics—i.e., the social lifeworld—and the characteristics of agent control architectures that such an environment warrants for agents to fulfil specific functions, such as generating believable dramatic plots.

Acknowledgments

The Austrian Research Institute for Artificial Intelligence is supported by the Austrian Federal Ministry for Education, Science and Culture and by the Austrian Federal Ministry for Transport, Innovation and Technology. This research is carried out within the Network of Excellence Humaine (Contract No. 507422) that is funded by the European Union's Sixth Framework Programme with support from the Austrian Funds for Research and Technology Promotion for Industry (FFF 808818/2970 KA/SA). This publication reflects only the author's views. The European Union is not liable for any use that may be made of the information contained herein.

References

URLs mentioned in footnotes or references have last been visited on 2005-01-12.

P. Agre and I. Horswill. Lifeworld analysis. *Journal of Artificial Intelligence Research*, 6:111–145, 1997.

M. Aubé. A commitment theory of emotions. In L. Cañamero, editor, *Emotional and Intelligent: The Tangled Knot of Cognition, Proc. of 1998 AAAI Fall Symposium, Orlando, FL, USA*, pages 13–18, 1998.

J. Bargh and T. Chartrand. The unbearable automaticity of being. *American Psychologist*, 54(7):462–479, 1999.

M. E. Bratman, D. J. Israel, and M. E. Pollack. Plans and resource-bounded practical reasoning. *Computational Intelligence*, 4(4):349–355, 1988.

M. Cavazza, F. Charles, and S. J. Mead. Character-based interactive storytelling. *IEEE Intelligent Systems*, special issue on AI in Interactive Entertainment 17(4):17–24, 2002.

F. Charles and M. Cavazza. Exploring the scalability of character-based storytelling. In N. e. a. Jennings, editor, *Proceedings of the third International Joint conference on Autonomous agents and multiagent systems (AAMAS'04), July 19-23, 2004, New York City, NY, USA, IEEE Computer Society*, volume 2, pages 872–879, 2004.

W. Clancey. *Conceptual Coordination: How the Mind Orders Experience in Time*. Lawrence Erlbaum Associates, Mahwah, New Jersey, London, 1999.

L. Egri. *The Art of Dramatic Writing*. Touchstone Book, New York, 1946.

C. Elliott, J. Brzezinski, S. Sheth, and R. Salvatoriello. Story-morphing in the affective reasoning paradigm: generating stories semi-automatically for use with "emotionally intelligent" multimedia agents. In K. P. Sycara and M. Wooldridge, editors, *Proceedings of the second international conference on Autonomous agents (Agents98) St.Paul, MN, USA*, pages 181–188. ACM Press, New York, 1998.

C. B. Frankel. Toward the nature of animation: An architectural approach. In *Proceedings, AISB 2002, Society for the Study of Artificial Intelligence and the Simulation of Behaviour*, London, England, April 2002.

N. Frijda. *The Emotions*. Cambridge University Press, Paris, Editions de la Maison des Sciences de l'Homme, 1986.

J. Gratch and S. Marsella. A domain-independent framework for modeling emotion. *Cognitive Systems Research*, 5(4):269–306, 2004.

M. J. Huber. JAM: a BDI-theoretic mobile agent architecture. In O. Etzioni, J. Mueller, and J. Bradshaw, editors, *Proceedings of the third annual conference on Autonomous Agents*, pages 236–243. ACM Press, Seattle, WA, USA, 1999.

M. J. Huber. *JAM Agents in a Nutshell*. Intelligent Reasoning Systems, Oceanside, CA, USA, Nov 2001. URL `http://www.marcush.net/IRS/Jam/Jam-man-01Nov01-draft.htm`.

D. Isla, R. Burke, M. Downie, and B. Blumberg. A layered brain architecture for synthetic creatures. In B. Nebel, editor, *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI 2001), Seattle WA USA, August 4-10*, pages 1051–1058. Morgan Kaufmann, San Francisco, CA, USA, 2001.

T. Kemper. Sociological models in the explanation of emotions. In M. Lewis and J. Haviland, editors, *Handbook of Emotions*, pages 41–52. Guilford Press, New York/London, 1993.

I. Machado, P. Brna, and A. Paiva. Learning by playing—supporting and guiding story-creation activities. In J. Moore, C. Redfield, and W. Johnson, editors, *International Conference on AI in Education*, volume 68 of *Frontiers in Artificial Intelligence and Applications*. IOS Press, Amsterdam, 2001. URL `http://gaips.inesc.pt/gaips/shared/docs/Machado01LearningByPlaying.pdf`.

B. Magerko, J. Laird, M. Assanie, A. Kerfoot, and D. Stokes. AI characters and directors for interactive computer games. In D. McGuinness and G. Ferguson, editors, *Proceedings of the 19th National Conference on Artificial Intelligence, 16th Conference on Innovative Applications of Artificial Intelligence (AAAI-04), July 25-29, 2004, San Jose, CA, USA*, pages 877–883. AAAI Press, Menlo Park, CA, 2004.

R. Marinier and J. Laird. Toward a comprehensive computational model of emotions and feelings. In M. Lovett, C. Schunn, C. Lebiere, and P. Munro, editors, *6th International Conference on Cognitive Modelling (ICCM2004): Integrating Models, July 30-August 1, 2004, Carnegie Mellon University, Pittsburgh, PA, USA, Learning Research and Development Center, University of Pittsburgh*, 2004.

S. C. Marsella, W. L. Johnson, and C. LaBore. Interactive pedagogical drama. In C. Sierra, M. Gini, and J. Rosenschein, editors, *Proceedings of the 4th International Conference on Autonomous Agents, Barcelona Spain, June 3-7 2000*, pages 301–308. ACM Press, New York, NY, USA, 2000.

M. Mateas and A. Stern. Architecture, authorial idioms and early observations of the interactive drama façade. Technical Report CMU-CS-02-198, School of Computer Science, Carnegie Mellon University, 2002.

A. Ortony. On making believable emotional agents believable. In R. Trappl, P. Petta, and S. Payr, editors, *Emotions in Humans and Artifacts*, pages 189–212. MIT Press, Cambridge MA/London UK, 2003.

P. Petta. The role of emotions in a tractable architecture for situated cognizers. In R. Trappl, P. Petta, and S. Payr, editors, *Emotions in Humans and Artifacts*, pages 251–287. MIT Press, Cambridge MA/London UK, 2003.

S. Rank. Affective acting: An appraisal-based architecture for agents as actors. M.S. thesis, Vienna University of Technology, Vienna, Austria, EU, 2004.

R. Reisenzein. Appraisal processes conceptualized from a schema-theoretic perspective: Contributions to a process analysis of emotions. In (Scherer et al., 2001), pages 187–201.

K. R. Scherer, A. Schorr, and T. Johnstone, editors. *Appraisal Processes in Emotion: Theory, Methods, Research*. Oxford University Press, Oxford/New York, 2001.

A. Sloman and M. Scheutz. A framework for comparing agent architectures. In *Proceedings UKCI 02: UK Workshop on Computational Intelligence, Sept 2002, Birmingham, UK*, 2002.

A. Staller and P. Petta. Introducing emotions into the computational study of social norms: A first evaluation. *Journal of Artificial Societies and Social Simulation*, 4(1), 2001.

C. Vogler. *The Writer's Journey: Mythic Structure for Storytellers and Screenwriters*. Boxtree, London, 1996.

Symbolic Objects and Symbolic Behaviors: Cognitive Support for Emotion and Motivation in Rational Agents*

Antônio Carlos da Rocha Costa*
*Escola de Informática - UCPel
96.010-000 Pelotas, RS, Brazil
rocha@atlas.ucpel.tche.br

Paulo Luis Rosa Sousa†
†Curso de Mestrado em Saúde e Comportamento - UCPel
96.010-000 Pelotas, RS, Brazil
psousa@phoenix.ucpel.tche.br

Abstract

This paper concerns the possible roles cognition plays with respect to emotion and motivation in rational agents. The approach we introduce is built around the notion of *symbolic object*, an object that reminds the agent of (possibly many) emotionally charged situations in which the agent was involved, so that facing the object and remembering those situations may operate as a motivator for specific actions concerning the current situation.

1. A *symbolic object* is an object that reminds an agent of past situations the agent was involved in, and that the agent felt were emotionally charged, so that facing that object again and remembering those situations may operate as a motivator for specific actions concerning the current situation. Symbolic objects, and the symbolic behaviors introduced later, are proposed as cognitive supports for a mental mechanism able to explain emotionally-based motivated behavior. We present a tentative conceptual framework for the working of symbolic objects and symbolic behaviors, and indicate some possible applications in the study of rational agents.

2. We assume the following sets as primitives: Obj, a set of objects; $Ag \subseteq Obj$, a set of agents; Sit, a set of situations that agents can recognize, concerning objects and other agents. We leave open the details about the elements of such sets. We just require that Sit, even if implicitly, encompasses a temporal structure. Situations are related to objects and agents by the following functions: $obj : Sit \rightarrow \wp(Obj)$, the function that indicates the set $obj(s)$ of objects that participate in situation s; $ag : Sit \rightarrow \wp(Ag)$, the specialization of obj to the set Ag, so that $ag(s)$ is the set of agents that participate in s.

Let relation $\Rrightarrow \subseteq Obj \times Ag \times Sit \times Sit$ be the *symbolization relation*, by which an object o in situation s symbolizes situation t to agent a, denoted $s \Rrightarrow_a^o t$. Then, we say that an object o is a *symbolic object* for an agent a in situation s if and only if there is a situation $t \neq s$ such that $s \Rrightarrow_a^o t$. An object $o \in obj(s)$ is said to *remind* an agent a of a situation t, in situation s, if and only if $s \Rrightarrow_a^o t$ and $o \in obj(t)$. Thus, reminding is conceived as justified symbolization.

3. Agents recognize situations by recognizing the objects (and agents) that participate in them, but they do that under the influence of symbolic objects, and the emotional charges that they convey. Thus, agents may be unable to tell situations completely apart. Situations that cannot be completely told apart by a given agent, due to the symbolic character of some objects participating in them, are said to be *symbolically similar* to that agent.

Let $\approx \subseteq Ag \times Sit \times Sit \times Obj$ be the *symbolic similarity* relation, by which an agent a considers situation s similar to situation t by virtue of the presence in s and t of the object o. Denote $s \approx_a^o t$ such relationship. We require that $\approx_a^o$ be reflexive and symmetric. Let $EmCh$ be a set of values, called *emotional charges*, and $emch : Ag \times Sit \rightarrow EmCh$ be the function that gives the emotional charge of any situation s for any agent a, denoted $emch_a(s)$. We require that: there is a *null* emotional charge, denoted $0 \in EmCh$; there is a relation $\leq$ between emotional charges, so that $(EmCh, \leq)$ is a partially ordered set. We also require that if $a \in ag(s)$ and $s \approx_a^o t$ then $emch_a(t) \neq 0$, so that only situations where agents are not involved, neither directly nor by association, are without emotional charges for them.

4. We consider agent behaviors to be coordinated sets of actions, performed by agents in the pursuit of *goals*. That is, we take into account only *goal-directed behavior*.

Let Beh be the set of all possible behaviors that agents may have. Let $beh : Ag \times Sit \rightarrow Beh$ the function that determines the behavior each agent a has in each situation s, denoted $beh_a(s)$. Let $Goal \subseteq Sit$ be the set of all goals that agents may have. Let $gl : Ag \times Sit \rightarrow \wp(Goal)$ be the function that determines the goals an agent may have at a given situation, denoted $gl_a(s)$. Let $\rightarrow \subseteq Sit \times Beh \times Sit$ be the relation that determines, for a situation s, that an agent behavior b may attain a situation t, denoted $s \stackrel{b}{\rightarrow} t$.

*This work was partially supported by CNPq and FAPERGS.

Then, for any rational agent a, goal t, and situation s: $t \in gl_a(s)$ if and only if $beh_a(s) = b$ and $s \xrightarrow{b} t$.

5. In the set of possible situations in which an agent may find itself, some will be felt as *emotionally acceptable*, others as *emotionally unacceptable*. The favorable situations are those situations s whose emotional charges are positive ($emch_a(s) > 0$), while the unfavorable ones are those with negative emotional charges ($emch_a(s) < 0$).

6. A goal t is an *emotional motivation* for a behavior b of an agent a in situation s, if and only if $emch_a(t) > 0$ and $s \xrightarrow{b} t$. An *emotional goal* is a goal generated only for the sake of the emotional charge it carries with it. A behavior is an *emotionally defensive behavior* in a situation if the situation is emotionally unacceptable to the agent and the behavior is performed only for the sake of taking the agent to another, emotionally acceptable situation. An agent *acts in a purely emotional way* when it changes from one behavior to another just for the sake of the bigger emotional charge implied by the new behavior.

7. An *interaction* between two agents is a coordinated performance of behaviors, so that rational relationships of dependence and causality are established among the behaviors that constitute the interaction. The set of rational relationships is called the *rational structure* of the interaction. During an interaction, agents change their behaviors according to the changes in the situations in which they find themselves. Thus, through changes in situations, behaviors of an agent cause changes in the behavior of the other agent. An interaction is said to be an *emotional interaction* if the emotional charges of the situations involved in the interaction play a role in the causation of the behaviors of the agents, in superposition to the causation determined by the interaction's rational structure.

Let a and a' be two interacting agents. Let $b_i = beh_a(s_i)$ and $b'_i = beh_{a'}(s_i)$ be the successive behaviors of a and a' during the successive situations $s_1, \ldots, s_n$. A *stage* of the interaction is any pair of behaviors (b_i, b'_i) performed by the agents that participate in the interaction. The interaction stage (b_i, b'_i) is said to be *emotionally favorable* for agent a (resp., a') in situation s_i if and only if $emch_a(s_i) < 0$ (resp., $emch_{a'}(s_i) < 0$), $emch_a(s_{i+i}) > 0$ (resp., $emch_{a'}(s_{i+i}) > 0$), and $s_i \xrightarrow{(b_i, b'_i)} s_{i+1}$.

8. Two agents are said to be *friendly to each other* in a situation s, which is emotionally unacceptable for one of them, if and only if they perform an interaction that leads them to a resulting situation which is emotionally acceptable for both. Only rational agents capable of emotionally motivated behaviors can be friendly to each other.

A sequence of stages $(b_j, b'_j), \ldots, (b_k, b'_k)$, starting in a situation s_j and ending in situation s_k, is said to be a *friendly interaction* in s_j if and only if $emch_a(s_j) < 0$ or $emch_{a'}(s_j) < 0$, and the both $emch_a(s_k) > 0$ and $emch_{a'}(s_k) > 0$.

9. We extend the relation of symbolization to encompass behaviors. Let $\Rrightarrow \subseteq Ag \times (Obj \cup Beh) \times Sit \times Sit$ be the symbolization relation extended to encompass symbolic behaviors, as well as symbolic objects. A *symbolic behavior* is one that is perceived symbolically by an agent: a behavior $b' = beh_{a'}(s)$ of an agent a' in situation s is a *symbolic behavior* for an agent a in s if and only if there exists a situation $t \neq s$ such that $s \Rrightarrow_a^{b'} t$.

10. A situation s is an *emotionally misleading situation* for agents a and a' if and only if there is an element x (either an object $o \in obj(s)$, or a behavior $b = beh_a(s)$ or $b' = beh_{a'}(s)$) that symbolizes different situations for the two agents: $s \Rrightarrow_a^x t$, $s \Rrightarrow_{a'}^x t'$, and $t \neq t'$. If s is a stage in an interaction and it happens that either a or a' (or for both) emotionally react to s, it is possible that such reactions, due to their lack of rational connection to the rest of the interaction, make the interaction depart from its rational structure. We call *emotionally misleading interaction* any interaction that has at least one emotionally misleading stage.

Conclusion. This paper sketched in a tentative way a conceptual framework for symbolic objects and symbolic behaviors as cognitive supports for emotions and motivations. We think that the elaboration of this framework may further the study of rational agents in at least two directions: it may help agents to reason adequately about their own emotional behaviors, and about those of their (human and artificial) partners; and it may help artificial agents to better simulate human agents.

The few ideas presented here are certainly not enough to support such aims, and many main and auxiliary ideas are still missing; e.g., a relation of similarity between objects (and between behaviors), so that the symbolization relation may be based on similar, not necessarily equal, objects (or behaviors), avoiding the requirement that objects (and behaviors) be present in exactly the same form, in different situations, to become symbolic.

Finally, we note that our approach originated from a critical view of the area of AI by the first author, a critical view of the theory and practice of Psychoanalysis by the second author, and from our joint idea that the time is ripe for the development of useful formal theories based on Freud's work (see, e.g., Freud (1976, 1989)).

References

Sigmund Freud. *A Project for a Scientific Psychology*. Norton, New York, 1976. (Written in 1895).

Sigmund Freud. *An Outline of Psycho-Analysis*. Norton, New York, 1989. (Written in 1938-1939).

An Affective Model of Action Selection for Virtual Humans

Etienne de Sevin and Daniel Thalmann
Swiss Federal Institute of Technology
Virtual Reality Lab VRLab
CH-1015 Lausanne Switzerland
{etienne.desevin, daniel.thalmann}@epfl.ch

Abstract

The goal of our work aims at implementing progressively an action selection affective model for virtual humans that should be in the end autonomous, adaptive and sociable. Affect, traditionally distinguished from "cold" cognition, regroups emotions and motivations which are highly intertwined. We present a bottom-up approach by implementing first a motivational model of action selection to obtain motivationally autonomous virtual humans. For the adaptability of virtual humans and completeness of our affective model of action selection, we will define the interactions between motivations and emotions in order to integrate an emotional layer. In order to understand how they affect decision making in virtual humans, the motivations should represent more quantitative aspect of the decision making whereas emotions should be more qualitative one.

1 Introduction

One of the main problem to solve, when a motivational decision making for individual virtual humans is designed, is the action selection problem: "how to choose the appropriate behavior at each point in time so as to work towards the satisfaction of the current goal (its most urgent need), paying attention at the same time to the demands and opportunities coming from the environment, and without neglecting, in the long term, the satisfaction of the other active needs" (Cañamero, 2000).

In a bottom-up approach, we decide to implement first a motivational model of action selection because motivations are directly implied in the goal-oriented behaviours. Next we will add an emotional layer for the flexibility and the realism of the behaviours. The emotions stay longer in time than the motivations which need to be satisfied rapidly and can modify and modulate motivations according to Frijda (1995): "emotions alert us to unexpected threats, interruptions, and opportunities".

In this paper, we describe first our motivational model of action selection for virtual humans with his functionalities for the flexibility and the coherence of the decision making. We created a simulated environment for testing the model in real-time. Finally we explain how an emotion layer could be added to obtain an affective model of action selection.

2 The motivational model of action selection

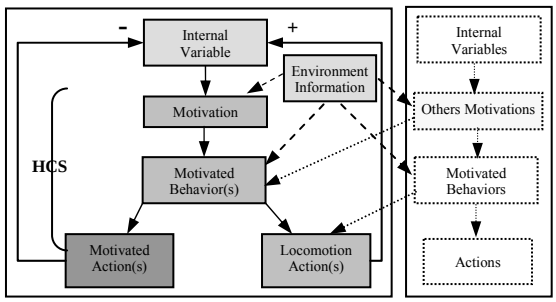


Figure 1: A hierarchical decision graph for one motivation connecting with others decision graphs

Our model is based on hierarchical classifier systems (Donnart and Meyer, 1994) (HCS, one per motivation) working in parallel to obtain goal-oriented behaviors for virtual humans. For the adaptability and reactivity of virtual humans in decision making, the HCS are associated with the functionalities of a free flow hierarchy (Tyrrell, 1993) such as compromise and opportunist behaviors. This model contains four levels per motivation:

- *Internal variables* represent the internal state of the virtual human and evolve according to the effects of actions.

- *Motivations* correspond to a "subjective evaluation" of the internal variables and environment information due to a threshold system and a hysteresis.

- *Motivated behaviors* represent sequences of locomotion actions, generated thanks to the hierarchical classifier system, to reach locations where the virtual human should go to satisfy motivations.

- *Actions* are separated into two types. *Locomotion actions* are only used for moving the virtual human to a specific place, where *motivated actions* can satisfy one or several motivations. Both have a retro-action on internal variable(s). Locomotion actions increase them, whereas motivated actions decrease them.

The motivational model is composed of many hierarchical classifier systems running in parallel. The number of motivations is not limited. Selection of the most activated node is not carried out at each layer, as in classical hierarchy, but only in the end, as in a free flow hierarchy (the action layer). Finally the action chosen is the most activated permitting flexibility and reactivity in decision making of virtual human.

2.1 Evaluation of motivations

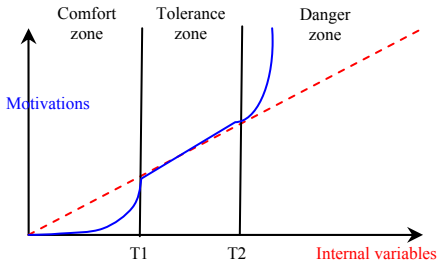


Figure 2: "Subjective" evaluation of one motivation from the values of the internal variable.

The "subjective evaluation" of motivations corresponds to a non-linear model of motivation evolution. A threshold system, specific to each motivation, reduces or enhances the motivation values, according to the internal variable values. This can be assimilated with levels of attention which limit and select information to reduce the complexity of the decision making task (Bryson, 2002). It helps the action selection mechanism to choose the most appropriate behavior at any time.

$$\begin{cases} M = T_1 e^{(i-T_1)^2} & \text{if } i < T_1 \\ M = i & \text{if } T_1 \leq i \geq T_2 \\ M = \frac{i}{(1-i))^2} & \text{if } i > T_2 \end{cases}$$

where M is the motivation value, T_1 the first threshold and i the internal variable

If the internal variable lies beneath the threshold T1 (comfort zone), the virtual human doesn't take the motivation into account. If the internal variable is beyond the second threshold T2 (danger zone), the value of the motivation is amplified in comparison with the internal variable. In this case, the corresponding action has more chances to be chosen by the action selection mechanism, to decrease the internal variable.

Moreover a hysteresis has been implemented, specific to each motivation, to keep at each step a portion of the motivation from the previous iteration, therefore permitting the persistence of motivated actions:

$$M_t = (1-\alpha)\cdot M_{t-1} + \alpha(M + e_t)$$

where M_t is the motivation value at the current time, M the "subjective" evaluation of the motivation, e_t the environment variable and α the hysteresis value with $0 \leq \alpha \leq 1$.

The hyteresis maintain the activity of the motivations and the corresponding motivated actions for a while, even though the activity of internal variables decreases. Indeed, the chosen action must remain the most activated until the internal variables have returned within their comfort zone. The hysteresis limits the risk of action selection oscillations between motivations and permits the persistence of motivated actions and the coherence in decision making.

2.2 Behavioral planner

To reach the specific locations where the virtual human can satisfy his motivations, goal-oriented behaviors (sequences of locomotion actions) need to be generated, according to environment information and internal context of the hierarchical classifier system. It can be use also for the complex actions as cooking which need to follow order in the sequence of actions. Moreover a learning or evolution process can be implemented thanks to weights of classifiers to optimize behaviors.

Time steps	t0	t1	t2	t3	t4	t5	t6
Environment information	known food location, but remote				Near food	Food near mouth	No food
Internal context (Message List)		hunger					
			reach food location				
Actions				Go to food	Take food	Eat	
Activated rules		R0	R1	R2	R3	R4	

Table 1: Simple example for generating a sequence of action using a hierarchical classifier system.

In the example (table 1), hunger is the highest motivation and must remain so until the nutritional state is returned within the comfort zone. The behavioral sequence of actions for eating needs two internal classifiers (modifying internal context):

R0: if known food location and the nutritional state is high, then hunger.

R1: if known food is remote and hunger, then reach food location.

and three external classifiers (activating actions):

R2: if reach food location and known food is remote, then go to food.

R3: if near food and reach food location, then take food.

R4: if food near mouth and hunger, then eat.

Here, the virtual human should go to a known food location where he can satisfy his hunger, but needs to generate a sequence of locomotion actions to reach that place. In this case, two internal messages "hunger" and "reach food location" are added to the message list, thanks to the internal classifiers R0, then R1. They represent the internal state for the rules and remain until they are realized. To reach the known food location, two external classifiers (R2 and R3) activate locomotion actions (as many times as necessary). When the virtual human is near the food, the internal message "reach food location" is deleted from the message list and the last external classifier R4 activates the motivated action "eat", decreasing the nutritional state. Thereafter the internal message "hunger" is deleted from the message list, the food has been eaten and the nutritional state is returned within the comfort zone for a while.

2.3 Reactive architecture

As activity is propagated throughout the model according to the free flow hierarchy, and the choice is only made at the level of actions, the most activated action is chosen according to motivations and environment information. A greater flexibility and reactivity in the behavior, such as compromise and opportunist behaviors are then possible in spite of behavioral planner.

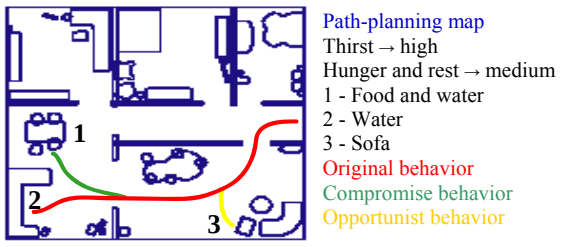


Figure 3: Compromise behavior (green): the virtual human goes where he can eat and drink instead of just drinking. Opportunist behavior (yellow): he stops to rest when he sees the sofa.

Compromise behaviors have more chances of being chosen by the action selection mechanism, since they can group activities coming from several motivations and can satisfy them at the same time. Opportunist behaviors occur when the virtual human perceives objects that can satisfy his motivations. These motivations are proportionally increased compared to the distance between objects and the virtual human. For these two beviahors, the propagated value in the model can be modified at two levels: at the motivations and motivated behaviors levels (see figure 1). If the current behaviour is exceeded, it is interrupted and a new sequence of locomotion actions is generated in order to reach the location where he can satisfy the new motivation.

3 Testing the model in a simulated environment

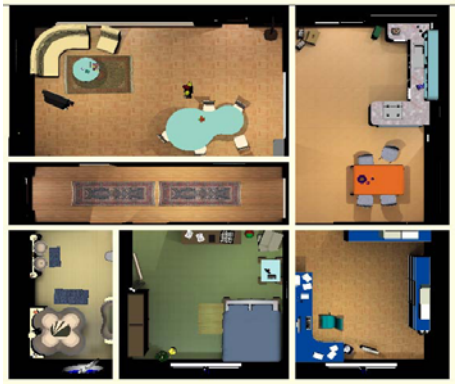


Figure 4: Top view of the simulated environment (apartment) in the 3D viewer.

We choose to simulate a virtual human in an apartment where he can "live" autonomously by perceiving his environment and satisfying several motivations. We arbitrarily define fourteen conflicting motivations (the number is not limited) that a human can have in this environment with their specific locations and the associated motivated actions.

motivations	locations	action
hunger	table	eat
thirst	sink, table	drink
toilet	toilet	satisfy
resting	sofa	rest
sleeping	bed	sleep
washing	bath	wash
cooking	oven	cook
cleaning	worktop, shelf	clean
reading	bookshelf	read
communicating	computer, phone	communicate
exercise	living, hall, room	do push-up
watering	plant	water
Watching (default)	sofa	watch TV
…	…	…

Table 2: all available motivations with associated actions and their locations.

At any time, the virtual human has to choose the most appropriate action to satisfy the highest motivation between conflicting ones according to environmental information. Then he goes to the specific place in the apartment where he can satisfy this motivation. Compromise behaviors are possible, for example the virtual human can drink and eat at the table. However he can perform different actions in the same place but not at the same time. The virtual human can also perform the same action in different places: for example clean at the worktop or at the shelf. Moreover he has a perception system to permit opportunist behaviors. The default action is watching television in the living room.

The users can add new motivations at the beginning, change all the parameters and monitor the different model level during the simulation.

4 Concluding remarks

The test application simulates in a 3D graphics engine (Ponder, 2003) a virtual human in an apartment, making decisions using the motivational model according to the motivations and the environment information. As a whole the action selection architecture doesn't oscillate between several motivations, managing the fourteen conflicting motivations, thanks to the hysteresis and the behavioral planner, and have also reactive behaviors such as compromise and opportunist behaviors. In the end the virtual human "lives" autonomously and adaptively in his apartment. Furthermore, the number of motivations in the model is not limited and can easily be extended.

The model has some limitations, though. For the time being, each motivation has the same importance in the decision-making process, although we know that some motivations are more important than others in real life. Here, the virtual human also always carries out the most activated action in any case. However, some actions should sometimes be delayed according to context.

5 Future work

Integrating emotions in the motivational model of action selection can reduce these limitations. First we plan to define the interactions between motivations, emotions and personalities to understand how they affect decision making in virtual humans. The main problem is to connect the emotional layer with the rest of the architecture. It could be made by a sort of synthetic physiology (Avila-Garcia and Cañamero, 2004). The motivations should be more quantitative aspect of the decision making whereas emotions should be the more qualitative one. The low level part of the architecture should be more automatic whereas the high level part should be more specified in real time by the users. The emotions will be independent from the motivations but influence them at the level of length, perception, activation and interruption. In the end we plan also to manage basic social interactions by adding another virtual humans in the apartment or/and by interacting directly with virtual reality devices in the next future.

References

Orlando Avila-Garcia and Lola Cañamero. Using Hormonal Feedback to Modulate Action Selection in a Competitive Scenario. *In Proceedings of the Eight Intl. Conf. on Simulation of Adaptive Behavior (SAB04)*: 243-252, Cambridge, MA: The MIT Press, 2004.

Johanna Bryson. Hierarchy and Sequence vs. Full Parallelism in Action Selection. *In Proceedings of the Sixth Intl.Conf. on Simulation of Adaptive Behavior (SAB00)*: 147-156, Cambridge, MA: The MIT Press, 2002.

Lola Cañamero. Designing Emotions for Activity Selection. *Dept. of Computer Science Technical Report* DAIMI PB 545, University of Aarhus, Denmark. 2000.

Jean-Yves. Donnart and Jean-Arcady Meyer. A hierarchical classifier system implementing a motivationally autonomous animat. *in the 3rd Int. Conf. on Simulation of Adaptive Behavior,* The MIT Press/Bradford Books, 1994.

Nico H. Frijda. Emotions in Robots. *In H.L. Roitlab and J.A. Meyer eds. Comparative Approaches to Cognitive Science*: 501-516, Cambridge, MA: The MIT Press, 1995.

Michal Ponder, Bruno Herbelin, Tom Molet, Sebastien Schertenlieb, Branislav Ulicny, George Papagiannakis, Nadia Magnenat-Thalmann, Daniel Thalmann. VHD++ Development Framework: Towards Extendible, Component Based VR/AR Simulation Engine Featuring Advanced Virtual Character Technologies. *in Comupter Graphics International (CGI)*, 2003.

Tobby Tyrrell. Computational Mechanisms for Action Selection, *in Centre for Cognitive Science*, Unievrsity of Edinburgh, 1993.

Integrating Domain-Independent Strategies into an Emotionally Sound Affective Framework for An Intelligent Learning Environment

Mohd Zaliman Yusoff
IDEAS lab, Department of Informatics,
School of Science &Technology,
University of Sussex, UK
m.z.yusoff@sussex.ac.uk.

Benedict du Boulay
IDEAS lab, Department of Informatics,
School of Science &Technology,
University of Sussex, UK
B.du-boulay@sussex.ac.uk

Abstract

This paper presents a framework that integrates domain-independent strategies into an emotionally sound affective (ESA) framework for an intelligent learning environment. The integration is an extension to current affective learning frameworks that consider only domain-dependent strategies to help student manage their emotional or affective states. It is hypothesised by helping students to manage their emotional or affective states, and hence, improve their performance in learning will improve.

Keywords: Domain-independent strategies, emotionally sound affective framework

1. Introduction

Despite the fact that emotions play an important role in learning, few attempts have been made to study emotions in Intelligent Tutoring Systems (ITS) though it is an area gaining increasing attention (e.g. Conati, 2002; del Soldato & du Boulay, 1995; Kort & Reilly, 2001; Lester et al., 1999a). Traditionally, affective learning frameworks use only domain-dependent strategies to help students manage their negative emotional or affective states. (e.g Conati, 2002; del Soldato & du Boulay,1995; Kort & Reilly, 2001; Lester et al., 1999a; Jaquas et al., 2002), for example by making the lesson easier if it is believed that the student needs some experience of success. However, emotion regulation theories have suggested that there are two strategies used to manage individual emotional or affective states: emotion-focused strategies, which are domain-independent and problem-focused strategies that are domain-dependent (Lazarus, 1991; Gross, 1999).

In this paper, we propose an emotionally sound affective (ESA) framework that integrates both domain-dependent and domain independent strategies. The ESA framework consists of two phases: 1) the appraisal phase, which attempts to appraise students' emotional state and 2) the reaction phase, which proposes to use adaptive strategies and activities, in order to help students manage their emotions (Yusoff, 2004).

The first phase of the ESA framework which appraises students' emotional states is introduced at two learning stages. The primary appraisal, which uses the PANAS questionnaire (Watson, Clerk & Tellegen, 1988) appraises students' emotional states at the beginning of a lesson. The primary appraisal establishes students' emotional states with regard to their personal beliefs and goal commitments. The secondary appraisal of this framework, on the other hand, appraises students' emotional states during the lesson. It uses students' reactions to two eliciting factors to appraise students' emotion. These eliciting factors are: the difficulty level of the lesson which is based on the nature of the lesson and the students' control over the lesson.

The students' Control over the lesson is modelled using student-computer interactions that are based on three methods: 1) by on-line communication with students during the interaction, 2) by monitoring students' request for help to complete a lesson and 3) students' self-reporting. In the ESA framework, asking for help, completing a lesson and, giving up are examples of the student-computer interactions. The intensity of the Control eliciting factor is determined by its three eliciting variables: Independence, Effort and Competence, that are derived from students' motivation modelling techniques in learning (e.g

del Soldato & du Boulay, 1995; De Vincente & Pain, 1999).

Independence is defined as the degree that students prefer to work without asking others for help. It has been widely used as an important parameter to detect students' affective states in affective learning environments (del Soldato & du Boulay, 1995; De Vincente & Pain, (1999); Jaques et al., 2003). In the ESA framework, Independence is modelled by the frequency of requests. A low request frequency corresponds to a high level of independence and high request frequency means a low level of independence.

Effort is another popular parameter used to detect students' affective states in an affective learning environment (e.g del Soldato & du Boulay, 1995; De Vincente & Pain, 1999). Effort is defined as the degree of engagement that students display to accomplish a task. In this framework, Effort is represented by the frequency of interactions between a student and the system, such as clicking on a mouse or pressing a key. A high number of interactions indicates a high level of effort, and a low number of interactions indicates otherwise.

Competence is the third variable that can influence the Control eliciting factor. It is a measure of the students' knowledge and skills to perform a lesson task proficiently. The framework represents Competence by a ratio of the number of errors to the number of attempts made to solve a problem. A low ratio corresponds to a high level of competence, and a high ratio implies otherwise.

Just as for the appraisal phase, the ESA framework implements the reaction phase at two learning stages: 1) at the beginning of a lesson and 2) during the lesson. Its main objective is to help students manage their emotions, especially after experiencing negative emotions, by using two underpinning strategies: domain independent or emotion-focused strategies and domain-dependent or problem-focused strategies.

The first strategies employed in the ESA framework in this reaction phase are the domain-dependent strategies. They help students by providing suitable suggestions and strategies that are adapted to the students' elicited emotional state and are based on the premise that students in a positive emotional state are more capable of mastering their lesson (Fredrickson, 1998).

In addition to the domain-dependent strategies, the domain independent or emotion-focused strategies are implemented to help student manage their emotions. Coping statements and relaxation exercises are examples of domain independent strategies. Statements such as "I can make things happen" are used to maintain students' happiness while statements like "I can see this problem from another perspective to make it seem more bearable" are used to reduce students' nervousness. Apart from coping statements, relaxation activities such as muscle and head exercises are employed to help students manage their emotions.

The focus of this paper is on the use of domain-independent strategies in the reaction phase of the ESA framework as a way to help students manage their negative emotional or affective states. Domain-independent strategies refer to strategies and techniques that are unrelated to the lesson domain. Coping statements and relaxation exercises are examples of domain-independent strategies. In contrast, traditional affective frameworks help students by adapting domain-dependent strategies to their emotional or affective states (e.g Conati, 2002; del Soldato & du Boulay,1995; Kort & Reilly, 2001; Lester et al., 1999a; Jaquas et al., 2003).

We postulate that the integration of domain-independent strategies into the ESA framework helps students to manage their emotional states better and hence improves their performance in learning. The complete flowchart the integration of both domain-independent and domain-dependent strategies is given in Figure 1.

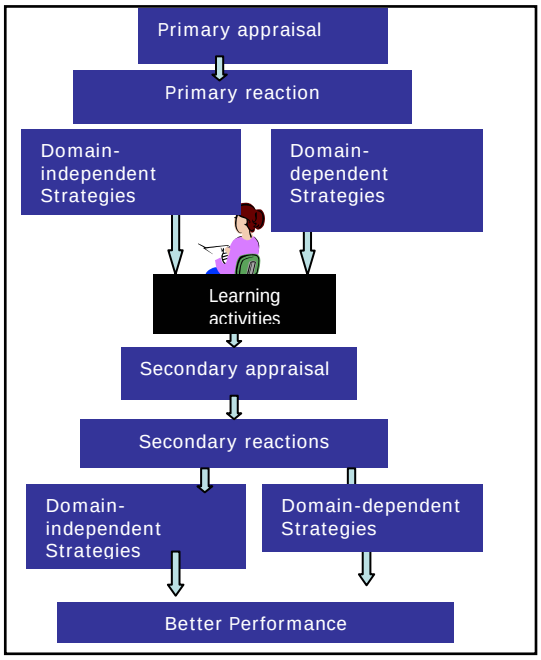


Figure 1: The flowchart of an emotionally sound affective (ESA) framework

2. Domain-independent strategies

To model the domain-independent strategies in the ESA framework, we refer to emotion regulation theories that are used to help individuals manage their emotional states (e.g Gross, 1999; Lazarus, 1991). Gross (1999) defines emotion regulation as a process by which individuals influence which emotions they have, when they have them, and how they experience and express these emotions. From Lazarus' (1991) viewpoint, emotion regulation consists of behaviour or cognitive responses or strategies that are designed to reduce, overcome, or tolerate the demands placed on the individual. These strategies are classified into two major categories: emotion-focused strategies and problem- focused strategies.

Emotion-focused strategies refer to thought or actions whose goal is to relieve the emotional impact of stress. There are apt to be mainly palliative in the sense that such strategies for coping do not actually alter the threatening or damaging conditions but make the person feel better. Examples are avoiding thinking about trouble, denying that anything is wrong, distancing or detaching oneself as in joking about what makes one feel distressed, or attempting to relax.

Problem-focused strategies, on the other hand, refer to efforts to improve the troubled person-environment relationship by changing things, for example, by seeking information about what to do, by holding back from impulsive and premature actions, and by confronting the person or persons responsible for one's difficulty

Therefore, we postulate that an emotionally sound affective framework must employ both domain-dependent and domain-independent in order to help students manage their negative emotional or affective states in learning.

3. Implementation of domain-independent strategies in ESA framework

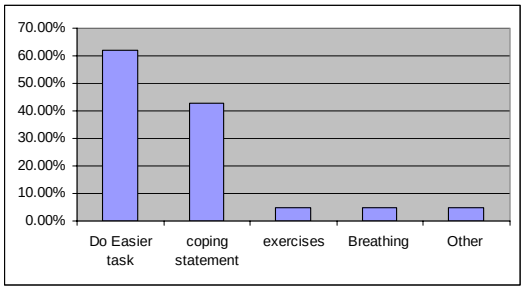
The emphasis of this framework is not to regulate student emotions completely, but more appropriately, to help students manage their emotions. Domain-independent strategies involve applying several strategies and techniques that are unrelated to lesson activities. Coping statements and relaxation exercises are examples of domain independent strategies. For example, for a nervous student, who has given up on a difficult lesson task, this framework suggests that he takes a deep breath several times, imagines a pleasurable and relaxing scene, and reads an effective coping statement such as "I won't let my sadness affect my performance" to reduce his nervousness. Apart from coping statements, relaxation activities such as muscle and head exercises are used to help students manage their emotions. As a result, students will feel better and, consequently help the student to learn better.

To implement the domain-independent efficiently in this framework, a general algorithm is being designed that combines several domain-independent strategies and is summarised as follows:

1. Getting loose (comfortable position)
2. Breathing exercises.
3. Doing muscles (head and eyes) relaxation exercises.
4. Reading coping statements.

An initial survey among 21 Sussex University students has indicated that besides domain dependent strategies, domain independent strategies are seen to be equally important in order to help them manage their negative emotions as shown in Table 1.

Table 1: The strategies preferred by Sussex University students in managing their negative emotion in learning.



4. Discussion

This paper emphasises more on the integration of domain-independent strategies into an affective learning framework which are derived from emotions regulations theories (Gross, 1999; Lazarus, 1991). Our initial finding, working with UK and Malaysian students, suggested that these domain independent strategies are helping them to manage their emotional states. By contrast, current affective learning frameworks use only domain-dependent strategies to help students manage their negative affective state. (e.g Conati, 2002; del Soldato & du Boulay,1995; Kort & Reilly, 2001; Lester et al., 1999a; Jaquas et al., 2002).

An empirical study to find more evidence of the efficiency of domain-independent strategies in laboratory environments without affecting their learning focus will be conducted as future work. Apart from finding empirical evidence of the domain-independent strategies efficiency in laboratory environments, cultural differences among students is another important issue to be explored in future. Initial work with UK and Malaysian students has shown that domain-independent strategies such as the use of coping statements are cultural dependent, and thus, indicated that it is important to establish which strategies are best suited in multi culture learning environments.

References

Conati, C. (2002). Probabilistic assessment of user's emotions in education games. Journal of Applied Artificial Intelligence, 16(Special Issue on managing cognition an Affect in HCI), 555-575.

De Vincente, A., & Pain, H. (1999). Motivation self-report in ITS. [*Proceeding of the Ninth World Conference on Artificial Intelligence in Education*]S. P. Lajoie & M. Vivet (Eds.), (pp. 651-653). Amsterdam: IOS Press.

del Soldato, T., & du Boulay, B. (1995). Implementation of motivational tactics in tutoring systems. *Journal of Artificial Intelligence in Education*, 4(6), 337-378.

Gross, J. J. (1999). Emotion and emotions regulation. In L. A. Pervin & O. P. John (Eds.), Handbook of Personality: Theory and Research (2nd Edition) (pp. 525-552). Guildford, New York.

Jaques, P., Andrade, A., Jung, J., Bordini, R., & Vicari, R. (20032). Towards user's emotion recognition: A case in an educational system. 2*nd International conference on Computational Intelligence, Robotics and Autonomous Systems (CIRAS)*, December 15-18, 2003, Singapore

Kort, B., & Reilly, R. (2001). An affective model of interplay between emotions and learning: reengineering educational pedagogy- building a learning companion . *In proceeding of IEEE International Conference on Advance Learning Technology.*

Lazarus, S. R. (1991). Emotion and Adaptation. Oxford U. Press.

Lester, J. C., Voerman, J. L., Towns, S. G., & Callaway, C. B. (1999a). Deictic believability: coordinating gesture locomotion and speech in lifelike pedagogical agents. *Applied Artificial Intelligence*, 13, 383-414.

Watson, D., Cleark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative Affect: The PANAS scales. *Journal of Personality & Social Psychology , 54*(6), 1063-1070.

Yusoff, M.Z (2004). An emotionally sound affective framework for Intelligent Tutoring System (ITS), The 17th White House Papers Graduate Research In Informatics at Sussex, University of Sussex, June 2004.