A Cognitive and Neural Model for Adaptive Emotion Reading by Mirroring Preparation States and Hebbian Learning*

Tibor Bosse[1], Zulfiqar A. Memon[1, 2], Jan Treur[1]

[1]Vrije Universiteit Amsterdam, Department of Artificial Intelligence
De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands

[2]Sukkur Institute of Business Administration (Sukkur IBA),
Airport Road Sukkur, Sindh, Pakistan

{tbosse, zamemon, treur}@few.vu.nl
http://www.few.vu.nl/~{tbosse, zamemon, treur}

Abstract

Two types of modelling approaches exist to reading an observed person's emotions: with or without making use of the observing person's own emotions. This paper focuses on an integrated approach that combines both types of approaches in an adaptive manner. The proposed models were inspired by recent advances in neurological context. Both a neural model and a more abstracted cognitive model are presented. In the first place emotion reading is modelled involving (preparatory) mirroring of body states of the observed person within the observing person. This involves a recursive body loop: a converging positive feedback loop based on reciprocal causation between preparations for body states and emotions felt. Here emotion reading involves the person's own body states and emotions in reading somebody else's emotions: first the same feeling is developed by mirroring, and after feeling the emotion, it is imputed to the other person. In the second place, as an extension an adaptive process is modelled based on Hebbian learning of a direct connection between a sensed stimulus concerning another agent's body state (e.g., face expression) and an emotion imputation state. After this Hebbian learning process the emotion is imputed to the other agent before it is actually felt, or even without it is felt. Both the mirroring and Hebbian learning processes first have been modelled at a neural level, and next, in a more abstracted form at a cognitive level. By means of an interpretation mapping the paper shows the relation between the obtained cognitive model and the neurological model. In addition to specifications of both models and the interpretation mapping, simulation results are shown, and automated verification of relevant emerging properties is discussed.

1 Introduction

From an evolutionary perspective, mindreading (or having a Theory of Mind) in humans and some other kinds of animals has developed for a number of aspects, for example, intention, attention, emotion, knowing; e.g., (Baron-Cohen, 1995; Bogdan, 1997; Dennett, 1987; Goldman, 2006; Goldman and Spirada, 2004; Malle, Moses, and Baldwin, 2001). Two philosophical perspectives on having a Theory of Mind are Simulation Theory and Theory Theory; cf. (Goldman, 2006). In the first perspective it is assumed that mindreading takes place by using the facilities involving the person's own cognitive states that are counterparts of the cognitive states attributed to the other person. For example, the state of feeling pain oneself is

* Parts of this paper are based on work presented at the 8th IEEE/WIC/ACM International Conference on Intelligent Agent Technology (Memon and Treur, 2008), the 9th International Conference on Cognitive Modelling (Bosse, Memon, and Treur, 2009a), the 31st Annual Conference of the Cognitive Science Society (Bosse, Memon, and Treur, 2009b), and the 12th International Conference on Principles of Practice in Multi-Agent Systems (Bosse, Memon, and Treur, 2009c).

used in the process to determine whether the other person has pain. The second perspective is based on reasoning using knowledge about relationships between cognitive states and observed behaviour. For example, in (Bosse, Memon, and Treur, 2007a, 2007b) mindreading concerning another person's beliefs, desires and intentions was addressed from a Theory Theory perspective, and in (Memon and Treur, 2008) mindreading of emotions was addressed from a Simulation Theory perspective, where a person's own emotions are involved in the process of reading the other person's emotions.

More and more neurological evidence supports the Simulation Theory perspective, in particular the recent discovery of *mirror neurons*: preparation neurons that are activated both when preparing for an action (including a change in body state) and when observing somebody else performing a similar action; e.g., (Rizzolatti and Sinigaglia, 2008; Pineda, 2009; Iacoboni, 2008). However, work as described in (Pantic and Rothkrantz, 1997, 2000) shows the feasibility of automated approaches to emotion recognition where the person's own emotions are not involved. This feasibility at least refers to the technical point of view, but leaves open the question of neurological plausibility.

The current paper shows how both perspectives can co-occur, both from a technical and neurological perspective. A unified view on emotion reading is presented, where on the one hand mechanisms are available to perform emotion reading by simulation involving the person's own emotions based on mirroring (in line with the Simulation Theory perspective), but on the other hand by *Hebbian learning* process a mechanism is developed where emotions are recognised without involving the person's own emotions (resulting in a model part of which is in line with the Theory Theory perspective). This unified view is illustrated by both a neural and a more abstracted cognitive model for adaptive emotion reading, and by showing the mapping of the cognitive model onto the neural model.

The two adaptive emotion reading models presented are based on three ingredients originating in the neurological area: a *recursive body loop* to generate emotional responses and feelings, the *mirrorring* function of preparation neurons, and *Hebbian learning*. By Damasio (1999, 2003) preparation neurons are attributed a crucial role in generating and feeling emotional responses. In particular, using a 'body loop' or 'as if body loop', a connection between such neurons and the feeling of emotions by sensing the person's own body state is obtained; see (Damasio, 1999, 2003) or the formalisation presented in (Bosse, Jonker and Treur, 2008). The concept of recursive body loop is used as one of the points of departure. This causal cycle through preparation and feeling states is triggered by a stimulus and after an indefinite number of rounds ends up in an equilibrium for both states. By Hebbian learning an adaptive model for emotion reading has been obtained, which is able to develop a shortcut in emotion recognition. The Hebbian learning creates a direct connection from the stimulus (e.g., an observed facial expression) to the imputed emotion, bypassing the body loop with the person's own emotional states. Some simulation results are discussed, and formally specified dynamic properties of adaptive and non-adaptive emotion reading are shown, and it is discussed how they were verified against simulation traces.

Within AI and Cognitive Science models are often designed as cognitive level models. Currently the amount of neurological knowledge is growing fast. One apparent way to exploit these neurological resources computationally is by designing neural level models. One might even be led to a conclusion that it is better to only design models at a neurological level, and totally give up to model at a cognitive level. Within AI and Cognitive Science, it is more and

more recognised that models can be more 'embodied' to obtain their grounding in (physical or neural) reality. Models describing a person's internal functioning as fully immersed in physical reality can be designed on the basis of modelling concepts that are appropriate to describe the relevant neural and biological concepts and their dynamics (e.g., Port and van Gelder, 1995). Such concepts can be directly used to specify a neural level model. However, in line with (Jonker, Treur, and Wijngaards, 2002), it is still possible to exploit such concepts and relations as discussed in neurological literature in a more abstracted form in a cognitive level model, using more abstract mental states. This is also in line with Bickle (1998, pp. 205-208), where he illustrates a similar perspective for the folk psychological account in relation to a neurobiological account of Hawkin and Kandel's (1984a,1984b) case. It is also possible to make models at both levels, and, in addition to specify precisely defined (reduction) relations between concepts used in a cognitive level model and concepts used in a neural level model. This paper shows how this can be done. Both cognitive level and neural level model for adaptive emotion reading are introduced, and by means of an interpretation mapping, a relation between these two models have been shown (e.g., Kim, 2005; Treur, 2010).

Summarising, this paper addresses two *main research questions*:

- How can emotion reading by a person be modelled taking the person's own emotional states into account, and how can this be integrated in an adaptive manner with emotion reading without taking into account the person's own emotional states?
- How can state of the art neurological knowledge be exploited in modelling these emotion reading processes; how can they be modelled at a neural level and how in a more abstracted form at a cognitive level, and how do the obtained models at these two levels relate to each other?

The first research question is the primary one. The second one is more a meta-question about modelling methods in the light of the large amount of neurological resources becoming available.

The structure of this paper is as follows. First, in Section 2 the principles behind the approach are briefly reviewed. In Section 3 the neural level model for adaptive emotion reading is introduced. Some simulation results are shown in Section 4. Next, in Section 5 the cognitive level model is described. Some simulation results are shown in Section 6. In Section 7 it is discussed how automated verification of a number of relevant emerging properties was applied. Section 8 shows the mapping of the cognitive level model onto the neural level model. The paper is concluded with a discussion in Section 9.

2. Principles Behind the Approach to Adaptive Emotion Reading

Three main ingredients of the neural model to generate emotional responses and feeling states for a given stimulus are:

(1) a recursive body loop (cf. Damasio, 1999, 2003)

(2) the notion of mirror neurons (cf. Rizzolatti and Sinigaglia, 2008; Pineda, 2009; Iacoboni, 2008) and by the Simulation Theory perspective on mindreading (cf. Goldman, 2006).

(3) a Hebbian learning principle for the adaptive mechanism incorporated in the model (cf. Hebb, 1949; Bi and Poo, 2001; Gerstner and Kistler, 2002; Wasserman, 1989).

These ingredients are briefly discussed below.

Recursive (as if) body loop

The models presented in this paper exploit the idea of a recursive 'body loop' or 'as if body loop', inspired by Damasio (1999, 2003).

> 'The changes related to body state are achieved by one of two mechanisms. One involves what I call the 'body loop'. It uses both humoral signals (chemical messages conveyed via the bloodstream) and neural signals (electrochemical messages conveyed via nerve pathways). As a result of both types of signal the body landscape is changed and is subsequently represented in somatosensory structures of the central nervous system, from the brain stem on up. The change in the representation of the body landscape can partly be achieved by another mechanism, which I call the 'as if body loop'. In this alternate mechanism, the representation of body-related changes is created directly in sensory body maps, under the control of other neural sites, for instance, the prefrontal cortices. It is 'as if' the body had really been changed but it was not.' (Damasio, 1999, p. 79-80)

For a body loop this roughly proceeds according to the following causal chain:

sensing a stimulus → sensory representation of a stimulus → preparation for bodily response → body state modification → sensing the body state → sensory representation of the body state → feeling the emotion

Alternatively, an 'as if body loop' uses a shortcut:

preparation for bodily response → sensory representation of the bodily response

The sensory representation of a modified body state is considered as the basis for feeling the emotion:

> 'As for the internal state of the organism in which the emotion is taking place, it has available both the emotion as neural object (the activation pattern at the induction sites) and the sensing of the consequences of the activation, a feeling, provided the resulting collection of neural patterns becomes images in mind.' (Damasio, 1999, p. 79).

A main idea used in the models introduced here is that the body loop (or as if body loop) is extended to a *recursive* (as if) body loop by assuming that in turn the preparation of the bodily response is also affected by the state of feeling the emotion (cf. Damasio, 2003):

> 'The brain has a direct means to respond to the object as feelings unfold because the object at the origin is inside the body, rather than external to it. The brain can act directly on the very object it is perceiving. It can do so by modifying the state of the object, or by altering the transmission of signals from it. The object at the origin on the one hand, and the brain map of that object on the other, can influence each other in a sort of reverberative process that is not to be found, for example, in the perception of an external object.' (…)
>
> 'In other words, feelings are not a passive perception or a flash in time, especially not in the case of feelings of joy and sorrow. For a while after an occasion of such feelings begins – for seconds or for minutes – there is a dynamic engagement of the body, almost certainly in a repeated fashion, and a subsequent dynamic variation of the perception. We perceive a series of transitions. We sense an interplay, a give and take.' (Damasio, 2003, pp. 91-92)

So, in addition to the causal chains described above, also a causal connection

feeling the emotion → preparation for bodily response

is assumed, which makes the two loops (body loop and as-if body loop) recursive.

The bodily response and the feeling are assigned a level or gradation, expressed by a number, which is assumed dynamic. The causal cycle is modelled as a positive feedback loop, triggered by the stimulus and converging to a certain level of feeling and body state. Here in each round of the cycle the next body state has a level that is affected by both the level of the stimulus and of the feeling state, and the next level of the feeling is based on the level of the body state. This implies a pattern of gradual generation (and extinction) of an emotion upon a stimulus.

Mirroring

When as a stimulus another person's face is taken, via a recursive body loop, gradually higher and higher activation levels of the person's own feeling state are generated. Indeed there is strong evidence that (already from an age of 1 hour) sensing somebody else's facial expression leads (within about 300 milliseconds) to preparing for and showing the same facial expression (Goldman and Sripada, 2004, pp. 129-130). This has been further supported from the neurological side by the recent discovery of *mirror neurons*: preparation neurons with a mirroring function; cf. (Rizzolatti, Fogassi, and Gallese, 2001; Wohlschlager and Bekkering, 2002; Kohler, Keysers, Umilta, Fogassi, Gallese, and Rizzolatti, 2002; Ferrari, Gallese, Rizzolatti and Sinigaglia, 2008; Pineda, 2009; Rizzolatti, and Fogassi, 2003; Rizzolatti, 2004; Rizzolatti and Craighero, 2004; Iacoboni, 2008).

Not only experiments with animals but also experiments with humans have provided much information, for example, fMRI data from experiments, single cell recordings with epileptic patients, and analysis of patiernts with specific forms of brain damage. Also upon observing facial expression mirror neuron activity is reported, for example, in (Dapretto, Davies, Pfeifer, et al., 2006, p. 949) it is found:

> 'This fMRI study shows that children with Autism Spectrum Disorder have reduced activity in mirror neuron areas during imitation and observation of facial emotional expressions. Furthermore, activity in mirror neuron areas correlates with severity of disease in autistic children.'

Mirror neurons have their function due to the embedding in the neural circuits they are part of. These neural circuits involve connections and loops with different parts of the cortex (parts of frontal, temporal and parietal lobe), but also to other areas such as insula and limbic system.

For example, Iacoboni (2005, p. 632) indicates:

> 'Mirror neurons have been found in the ventral premotor cortex (…) and in the rostral sector of the inferior parietal lobule (…). F5 and PF are anatomically interconnected (…); in addition, PF connects with the superior temporal sulcus (STS) (…). In the STS, there are higher-order visual neurons that respond to seeing the actions of others (…). Thus, in the macaque, there seems to be a circuitry composed of the STS, PF and F5 that codes the actions of others and seems to be able to map these actions onto the motor repertoire of the observer.'

Moreover, in (Carr , Iacoboni, Dubeau, Mazziotta, and Lenzi, 2003, p. 5498) it is stated:

> 'A recent fMRI study of the observation and imitation of facial emotional expressions has revealed a large-scale neural network that comprises the core circuitry for imitation (the mirror neuron system and the STS), the insula and the limbic system'

Hebbian learning

Hebbian learning is based on the principle that connected neurons that are frequently activated simultaneously strengthen their connecting synapse. The principle goes back to Hebb (1949), but has recently gained enhanced interest by more extensive empirical support (e.g., Bi and Poo, 2001), and more advanced mathematical formulations (e.g., Gerstner and Kistler, 2002). In the models a variant of this principle has been adopted to realise a strengthened direct connection between sensory representation of stimulus and imputation.

3 A Neural Model for Adaptive Emotion Reading

In this section the neural model made by adopting the principles discussed in Section 2 is presented. The neural model was specified both in MatLab and in the hybrid dynamical modelling language LEADSTO (Bosse, Jonker, Meij, and Treur, 2007). Within this language, the temporal relation a $\rightarrow\!\!\!\!\rightarrow$ b denotes that when a state property a occurs, then after a certain time delay (which for each relation instance can be specified as any positive real number), state property b will occur. In LEADSTO, both logical and numerical calculations can be specified, and a dedicated software environment is available to support specification and simulation; for more details see (Bosse, Jonker, Meij, and Treur, 2007).

3.1 The Neural Network Structure

The neural model for adaptive emotion reading introduced here refers to activation states of (groups of) neurons and the body. An overall picture of the network structure of this model is shown in Figure 1. In the network structure depicted in Figure 1 each node stands for a group of one or more neurons, or for an effector, sensor or body state. The nodes can be interpreted as explained in Table 1.

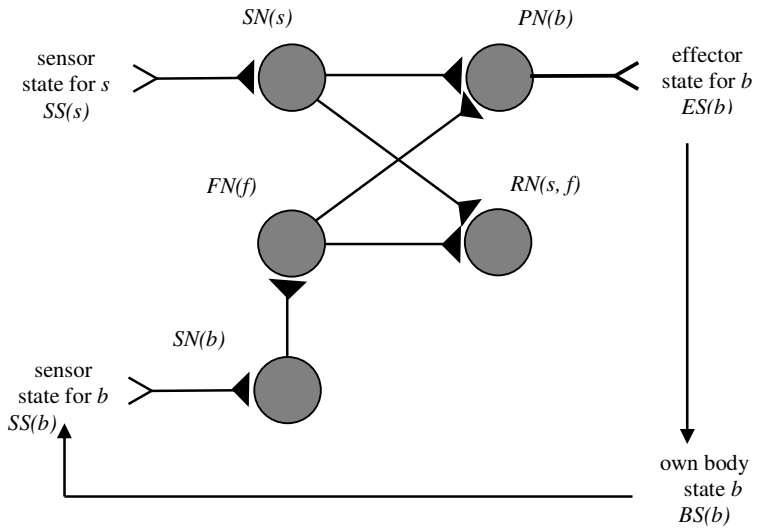

Figure 1: Network structure of the neural model for adaptive emotion reading

In the neural activation state of *RN(s, b)*, the experienced emotion *b* is related to the stimulus *s*, which triggers the emotion generation process. Note that to the extent that this neuron is related to *SN(s)*, it may be considered a basis for awareness of what causes the feeling *b*, which

may relate to what by Damasio (1999) is called a state of conscious feeling. This state that relates an emotion felt *b* to any triggering stimulus *s* can play an important role in the conscious attribution of the feeling to any stimulus *s*.

node nr	*denoted by*	*description*
0	*s*	stimulus; for example, another agent's body state b'
1	*SS(s)*	sensor state for stimulus *s*
2	*SN(s)*	sensory representation neuron for *s*
3	*PN(b)*	preparation neuron for the person's own body state *b*
4	*ES(b)*	effector state for the person's own body state *b*
5	*BS(b)*	person's own body state *b*
6	*SS(b)*	sensor state for the person's own body state *b*
7	*SN(b)*	sensory representation neuron for the person's own body state *b*
8	*FN(b)*	neuron for feeling state *b*
9	*RN(s, b)*	neuron representing that *s* induces feeling *b*

Table 1: Overview of the nodes involved

The neural model for emotion reading has been formally specified in LEADSTO. To this end the connections with their strengths were specified by:

connectedto(s, sensor_state(S), 1)
connectedto(sensor_state(S), SN(S), 1)
connectedto(FN(B), SN(S), PN(B), 0.5, 0.5)
connectedto(PN(B), effector_state(B), 1)
connectedto(effector_state(B), body_state(B), 1)
connectedto(body_state(B), sensor_state(B), 1)
connectedto(sensor_state(B), SN(B), 1)
connectedto(SN(B), FN(B), 1)
connectedto(FN(B), SN(S), RN(S, B), α, β)

3.2 Functioning of the Neural Model

According to the Simulation Theory perspective, an agent model for emotion reading should essentially be based on a neural model to generate the person's own emotions as induced by any stimulus *s*. The neural agent model introduced above has been specialised in a quite straightforward manner to enable emotion reading. The main step is that the stimulus *s* that triggers the emotional process, which until now was left open, is instantiated with the body state *b'* of another agent (for example a facial expression of another agent). Within the network in Figure 1 this leads (via activation of the sensory representation state *SN(b')*) to activation of the preparation state *PN(b)* where *b* is the person's own body state corresponding to the other agent's body state *b'*. This pattern shows how this preparation state *PN(b)* functions as a *mirror neuron*.

State variables and dynamic relations

To formally specify the functioning of the neural model, the mathematical concepts listed in Table 2 are used.

concept	*description*
N	set of nodes (as listed in Table 1); variables indicating elements of this set are i, j, k
N'	$N\backslash\{0\}$ the set of node numbers except the node for the stimulus s
$w_{ij}(t)$	strength of the connection from node i to node j at time t; this is taken 0 when no connection exists or when $i=j$
$y_i(t)$	activation level of node i at time t
$net_i(t)$	net input to node i at time t
g	function to determine activation level from net input
γ	change rate for activation level
η	learning rate for weights

Table 2: Mathematical concepts used

The function g can take different forms, varying from the identity function $g(v) = v$ for the linear case, to a discontinuous threshold (indicated by β) step function with $g(v) = 0$ for $v<\beta$ and $g(v)=1$ for $v\geq\beta$, or a continuous logistic threshold function based on $\frac{1}{1+e^{-\sigma(v-\tau)}}$ with steepness σ and threshold τ. For the connections between nodes of which at least one is not a neuron the connections have been made simple: weights 1 and g the identity function; so $w_{12} = w_{34} = w_{45} = w_{56} = w_{67} = 1$.

The activation levels are determined for step size Δt for all $i \in N'$ as follows:

$$net_i(t) = \Sigma_{j\in N}\ w_{ji}(t)\ y_j(t)$$
$$\Delta y_i(t) = \gamma(g(net_i(t)) - y_i(t))\ \Delta t$$

Note that for step size $\Delta t = 1$ and change rate $\gamma = 1$, the latter difference equation can be rewritten to

$$y_i(t+1) = g(net_i(t))$$

which is a wellknown formula in the literature addressing simulation with neural models.

The generic propagation rules for functioning of the neural model were specified in LEADSTO format as (corresponding to general neurological laws):

connectedto(X, Y, α) & activated(X, V) → activated(Y, α* V)
connectedto(X1, X2, Y, α, β) & activated(X1, V1) & activated(X2, V2) → activated(Y, α* V1+β *V2)

These temporal relations specify that propagation of activation levels takes place by multiplying them by the strength of the connection; for input from multiple connections they are added.

3.3 Hebbian Learning within the Neural Model

As a next step, the neural model for emotion reading is extended by a facility to strengthen the direct connection between the neuron $SN(s)$ for the sensory representation of the stimulus (the other agent's face expression) and the neuron $RN(s, f)$. A strengthening of this connection over time creates a different emotion reading process that in principle can bypass the generation of the person's own feeling.

Hebbian learning rule

The learning rule to achieve such an adaptation process is based on the Hebbian learning principle that connected neurons that are frequently activated simultaneously strengthen their

connecting synapse e.g., (Hebb, 1949; Bi and Poo, 2001; Gerstner and Kistler, 2002; Wasserman, 1989). The change in strength for the connection w_{ij} between nodes $i, j \in N$ is determined (for step size Δt) as follows; see also (Gerstner and Kistler, 2002, p. 406):

$$\Delta w_{ij}(t) = (\eta\, y_i(t) y_j(t)(1 - w_{ij}(t)) - \zeta\, w_{ij}(t))\, \Delta t$$

Here η is the learning rate, and ζ the extinction rate. Note that this Hebbian learning rule is applied only to those pairs of nodes $i, j \in N$ for which a connection already exists. In LEADSTO this rule was specified as:

connectedto(X1, X2, Y, W, β) & activated(X1, V1) & activated(Y, V2)
→ connectedto(X1, X2, Y, W+ηV1V2(1-W) - ζW, β)

By enabling the learning of a strengthened connection between sensory representation *SN(s)* of a stimulus and emotion imputation *RN(s, f)*, this neural model realizes that (after a learning phase) a person can perform emotion reading without taking the his or her own emotions into account. Part of this learnt model fits better in the Theory Theory perspective, than in the Simulation Theory perspective. A more extensive discussion about this debate is presented in Section 9.

In Appendix A an equilibrium analysis for the neural model can be found.

4. Example Simulations for the Neural Model

Based on the neural model specifications, a number of simulation traces have been generated, both within the LEADSTO environment and in MatLab. Time delays within the temporal LEADSTO relations were taken 1 time unit. Section 4.1 presents some simulation traces for the nonadaptive case, and Section 4.2 presents some traces for the adaptive case.

4.1 Nonadaptive example simulations of the neural model

An example simulation trace for the nonadaptive case is shown in Figure 2. The graphs show the values of the various activation levels (on the y-axis) over time (on the x-axis). Here it is shown that the recursive body loop results in an approximation of convergent activation levels for the states that relate to the emotion and the body state, among others. A simulation trace for emotion reading is obtained by instantiating stimulus `S` with the other person's face expression (indicated by `s = othersface(f)`), and instantiating body state `B` with the own face expression (indicated by `f`). Next, this trace is extended with a communication part, based on additional connections (see Figure 3):

connectedto(RN(S, B), PN(say(your emotion is B)), 1)
connectedto(PN(say(your emotion is B)), effector_state(say(your emotion is B)), 1)

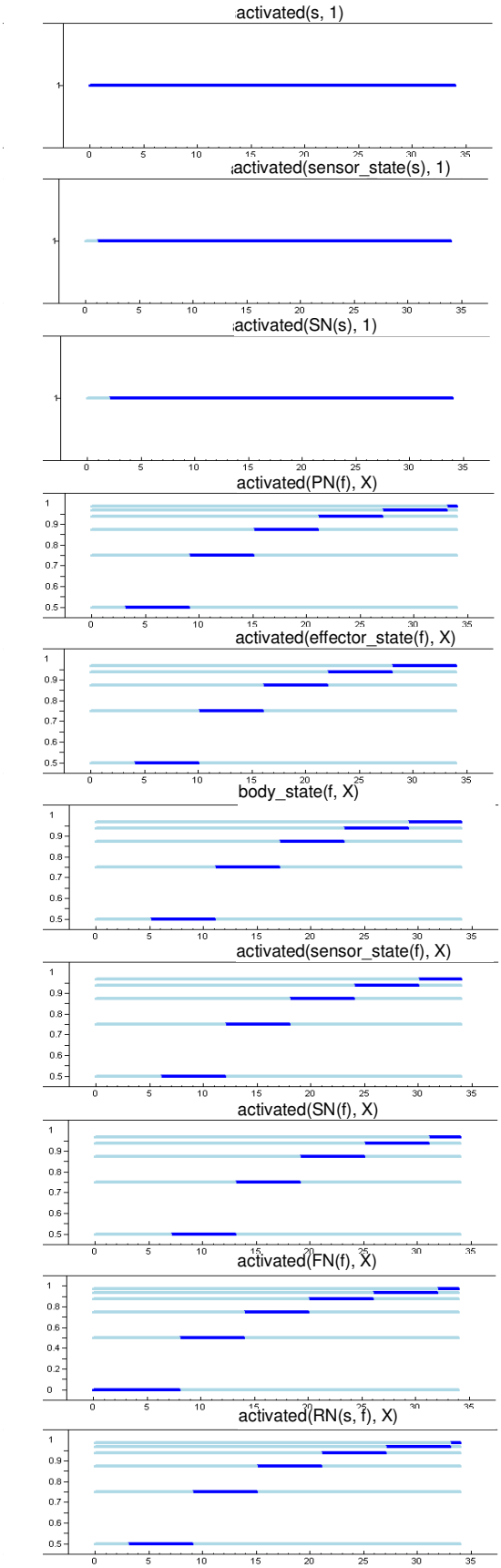

Figure 2: Example simulation trace for the neural model: nonadaptive case. These graphs show the values of the various activation levels (on the y-axis) from stimulus, sensing the stimulus, preparation for body state to sensing the body state over time (on the x-axis)

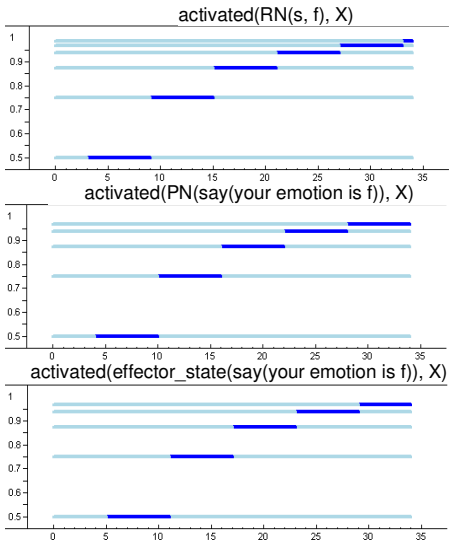

Figure 3: Example simulation trace for neural model with communication
These graphs show the values of the activation levels (on the y-axis) for imputation of the feeling, preparation of the communication and actual communication over time (on the x-axis)

Note that at time point 3 the neuron RN(s, f) has activation level *0.5*, which is not considered high enough to count as an indication of imputation. However, after time point 9 it gets an activation level of *0.75*. This is considered an appropriate indication for an imputation.

The numerical software environment Matlab has also been used to obtain simulation traces for the neural model described above. An example simulation trace that results from this neural model with the function g the identity function is shown in Figure 4. Here, time is on the horizontal axis, and the activation levels of three of the neurons *SN(s)*, *FN(f)*, and *RN(s,f)* are shown on the vertical axis. As shown in this picture, the sensory representation of a certain stimulus *s* quickly results in a feeling state *f*, and a representation that *s* induces *f*.

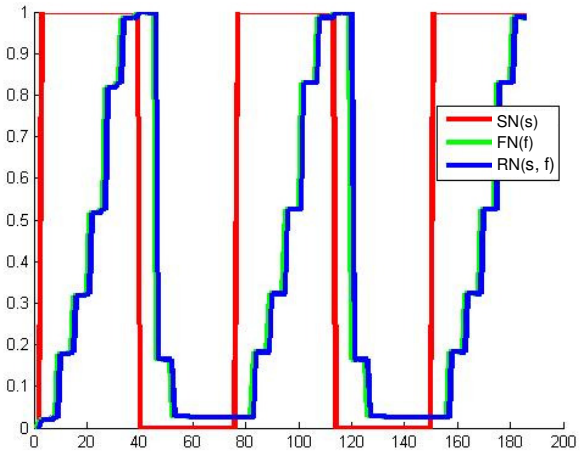

Figure 4: Example MatLab simulation for an agent performing non-adaptive emotion reading

When the stimulus s is not present anymore, the activations of *FN(f)* and *RN(s, f)* quickly decrease to *0*. The weight factors taken are: $w_{23} = w_{83} = w_{89} = 0.1$, $w_{78} = 0.5$ and $w_{29} = 0$. Moreover, $\gamma = 1$, and a logistic threshold function was used with threshold *0.1* and steepness *40*.

4.2 Adaptive example simulations of the neural model

Also a number of simulations have been performed for the neural agent model performing adaptive emotion reading; for an example, see Figure 5. As seen in this figure, the strength of the connection between *SN(s)* and *RN(s, f)* (indicated by *b* which is in fact w_{29}) is initially *0* (i.e., initially, when observing the other agent's face, the agent does not impute feeling to this). However, during an adaptation phase of two trials, the connection strength goes up as soon as the agent imputes feeling *f* to the target stimulus s (the observation of the other agent's face), in accordance with the temporal relationship described above.

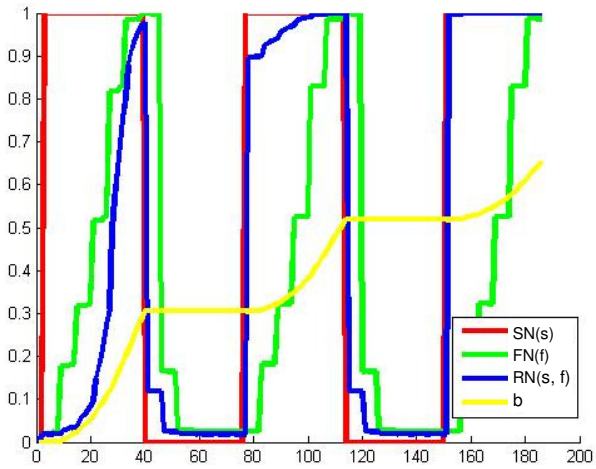

Figure 5: Example simulation for the neural model performing adaptive emotion reading

Note that, as in Figures 3 and 4, the activation values of other neurons gradually increase as the agent observes the stimulus, following the recursive body loop discussed. These values sharply decrease as the agent stops observing the stimulus as shown in Figures 4 and 5, e.g. from time point 40 to 76, from time point 112 to 148, and so on. Note that at these time points the strength of the connection between *SN(s)* and *RN(s, f)* (indicated by *b*) remains stable. After the adaptation phase, and with the imputation sensitivity at high, the agent imputes feeling *f* to the target stimulus directly after occurrence of the sensory representation of the stimulus, as shown in the third trial in Figure 5. Note here that even though the agent has adapted to impute feeling *f* to the target directly after the stimulus, the other state property values continue to increase in the third trial as the agent receives the stimulus; this is because the adaptation phase creates a connection between the sensory representation of the stimulus and emotion imputation without eliminating the recursive loop altogether.

The learning rate η used in the simulation shown in Figure 5 is *0.02*, the extinction rate was put on *0*. In Figure 6 a similar simulation is shown for a lower learning rate: *0.005*.

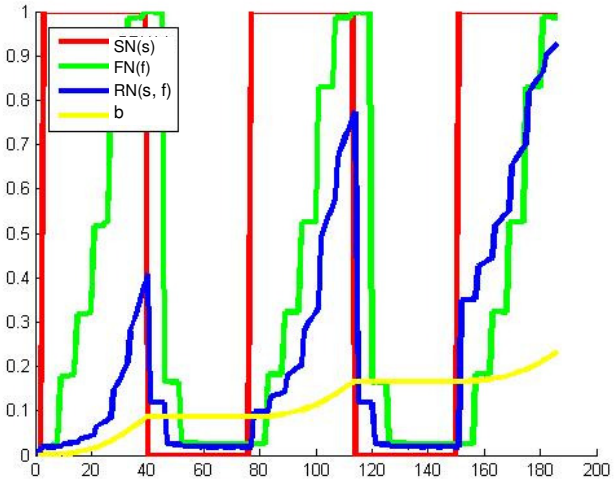

Figure 6: Adaptive emotion reading of the neural model with lower learning rate

5. A Cognitive Model for Adaptive Emotion Reading

The adaptive cognitive model to generate emotional responses and feeling states for a given stimulus was obtained by abstracting three main ingredients from neurological principles. More specifically, the following principles (also used as a basis for the neural model) as discussed in Section 2 were abstracted to a cognitive level:

(1) a *recursive body loop* for cognitive preparation states and feeling states (cf. Damasio, 1999, 2003),

(2) a *mirrorring function* of cognitive preparation states as inspired by the notion of mirror neurons (cf. Rizzolatti and Sinigaglia, 2008; Pineda, 2009; Iacoboni, 2008) and by the Simulation Theory perspective on mindreading (cf. Goldman, 2006),

(3) a cognitive-level *Hebbian learning principle* for adaptivity in the cognitive model (cf. Hebb, 1949; Bi and Poo, 2001; Gerstner and Kistler, 2002; Wasserman, 1989).

These ingredients are addressed at the cognitive level, respectively, in Section 5.1, 5.2, and 5.3. The description of the detailed cognitive model for emotion generation based on a recursive body loop is presented briefly in the sections below and specified in full in LEADSTO in Appendix B.

5.1 Recursive Body Loop

Figure 7 shows a graphical representation, where circles denote cognitive state properties and arrows denote temporal relationships. Here capitals are used for (assumed universally quantified) variables, and lower case letters for instances. In the figure it is assumed that b is a body state instance induced by stimulus instance s. The first two properties LP1 and LP2 describe the sensing process, and are assumed to apply for all instances of the variable *S*. Note that states here are binary.

LP1 Sensing a stimulus

If stimulus S occurs, then a sensor state for S will occur.

LP2 Generating a sensory representation of a stimulus

If a sensor state for S occurs, then a sensory representation for S will occur.

The third property LP3 only applies to a given specific stimulus instance *s* and a specific body state instance *b*. Here states have a certain level *V*: a real number in the interval [*0, 1*].

LP3 From sensory representation and emotion to preparation

If a sensory representation for *s* occurs and feeling *b* has level V,
then the preparation state for body state *b* will occur with level (1+V)/2.
If no sensory representation for *s* occurs and feeling *b* has level V,
then preparation state for body state *b* will occur with level V/2.

Here, it is assumed that the relative effects of both antecedents are the same. However, the formula (1+V)/2 can as well be replaced by the more generic formula w_1+w_2*V with weights w_1 and w_2. Such a variation also enables the modeller to distinguish different types of emotions (e.g., fear may develop faster than happiness). The properties LP4 to LP8 describe the general pattern of the body loop and are applicable to all instances of variable *B*.

LP4 From preparation to body modification

If preparation state for body state *B* occurs with level V, then the body state will express *B* with level V.

LP5 From body modification to modified body

If the body state is modified to express *B* with level V, then the body state will have expression *B* with level V.

LP6 Sensing a body state

If body state B with level V occurs, then body state B is sensed.

LP7 Generating a sensory representation of a body state

If body state B of level V is sensed, then a sensory representation for body state B with level V will occur.

LP8 From sensory representation of body state to feeling the emotion

If a sensory representation for body state *B* with level V occurs, then body state *B* is felt with level V.

Property LP9 describes the imputation and applies to all instances of variables *S* and *B*.

LP9 Imputation

If a certain body state *B* is felt, with level ≥ th, and a sensory representation for *S* occurs,
then emotion *B* will imputed to *S*.
Here, th is a (constant) threshold for imputation of emotion. In the simulations shown, th is assumed 0.95.

In the imputation state, the experienced emotion B is related to the stimulus S, which triggers the emotion generation process.

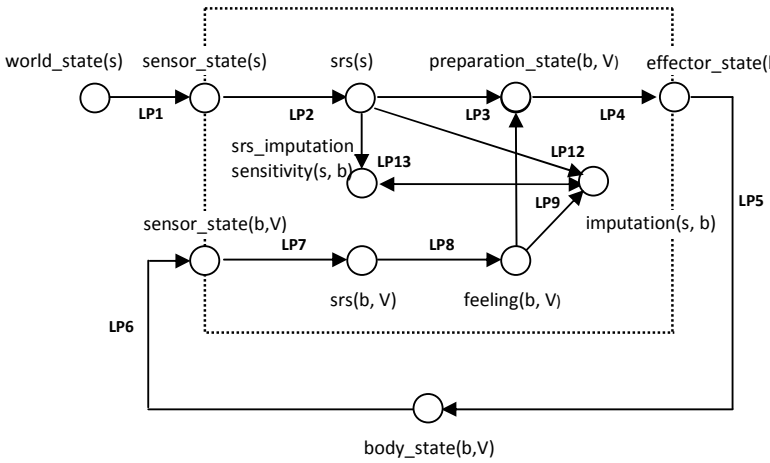

Figure 7: Cognitive model for adaptive emotion reading

Note that this state makes sense in general, for any type of stimulus S, as usually a person does not only feel an emotion, but also has an awareness of what causes an emotion; what by Damasio (1999) is called a state of conscious feeling also plays this role. This state that relates an emotion felt to a triggering stimulus plays an important role in the emotion reading process.

A recursive as-if body loop has been achieved by replacing the temporal relations LP4, LP5, LP6, LP7 by the following relation:

LP4* From preparation to sensory representation of body state
If preparation state for body state B occurs with level V,
then a sensory representation for body state B with level V will occur.

5.2 Emotion Reading by Mirrorring and Simulation

Based on the model for a recursive body loop, a model for emotion reading for the Simulation Theory perspective is obtained. Such a model for emotion reading uses the model to generate the person's own emotional responses and feelings to *simulate* the other person's process. The model presented above has been specialised in simple manner to enable emotion reading. The main step is to assume that for another person's body state that is observed (as a stimulus) a cognitive preparation state exists with a *mirrorring function*. This means that the stimulus s that triggers the emotional process is instantiated with the body state of another person, as was done in Section 3.2; to make it specific, a facial expression f of another person is considered; for example, s = othersface(f), and the body state instance b is face expression f.

For the sake of illustration, following the emotion imputation, a communication about it is prepared and performed. This extension is not essential for the emotion reading capability, but just shows an example of behaviour based on emotion reading.

LP10 Communication preparation
If emotion *B* is imputed to *S*, then a related communication is prepared

LP11 Communication
If a communication is prepared, then this communication will be performed.

5.3 Adaptivity of Emotion Reading Based on Hebbian Learning

This section extends the model presented above by a facility to learn a direct connection between the stimulus (the other person's body state) and the emotion imputation. An extra state is included that represents the sensitivity of how the emotion imputation depends on the sensory representation of the stimulus (the other face). At the cognitive level this can be expressed in qualitative or quantitative manners. If this sensitivity is qualified as 'high', the imputation will directly follow the sensory representation of the stimulus, as is expressed by the following temporal relationship.

LP12 Direct imputation
If the imputation sensitivity between *S* and *B* is high and a sensory representation for *S* occurs,
then emotion *B* will imputed to *S*.

The adaptation process itself and the persistence of the sensitivity level is described by the following two relationships.

LP13 Imputation sensitivity adaptation
If the imputation sensitivity from *S* to *B* is *W1* and a sensory representation for *S* occurs
and an imputation occurs for *B* to *S*, then the imputation sensitivity will become the value *W2* next to *W1*.

LP14 Imputation sensitivity persistence
If the imputation sensitivity is *W1* and no increase occurs,
then it will remain the same.

Note that the labels that represent the sensitivity levels may be elements of any linearly ordered set. Here, for simplicity, the set {low, medium, high} is taken, with relations next_value(low, medium) and next_value(medium, high). However, also other linearly ordered sets may be used, for example the set of real numbers between *0* and *1*.

By enabling the learning of a full connection between sensory representation of a stimulus and emotion imputation, this extended cognitive model entails that (after a learning phase) a person can perform emotion reading without taking his or her own emotions into account. As such, one could argue that part of this learnt model fits better in the Theory Theory perspective (not entirely but with certain aspects and in only a specific and simple form), than in the Simulation Theory perspective. A more extensive discussion about this debate is presented in Section 9.

6. Example Simulations for the Cognitive Model

Based on the cognitive model for adaptive emotion reading presented in Section 5, also a number of simulations have been performed; for an example, see Figure 8. Note that here the sensitivity values have been chosen as qualitative labels: *low, medium, high*. In this figure, the imputation sensitivity state has initial value set to low, represented by

srs_imputation_sensitivity(s, f, low)

in the upper part of Figure 8. In this part of the trace, a dark box on top of a line indicates that a state property is true at that time point, and a light box below the line indicates that the state property is false. The adaptation phase consists of two trials, where as soon as the person imputes emotion e to the target stimulus s (which is the observation of the other person's face), the imputation sensitivity level goes up, i.e., from low to medium to high, in accordance with the temporal relationship LP13 (see Section 5.3).

Note that the sensitivity state keeps its value in the adaptation phase until the person (again) imputes emotion f to the target, as described by the temporal relationship LP14, but retains its final value, i.e. high, after the adaptation phase of two trials. Moreover, note that in the lower part of Figure 8, the values of other state properties gradually increase as the person observes the stimulus, following the recursive body loop discussed in Section 5. These values sharply decrease as the person stops observing the stimulus, as described by the temporal relationship LP3 in Section 5.1. After the adaptation phase, and with the imputation sensitivity at high, the person imputes emotion f to the target stimulus directly after occurrence of the sensory representation of the stimulus, as shown in the third trial in the upper part of Figure 8. Again, note that, even though the person has adapted to impute emotion f to the target directly after the stimulus, the other state property values continue to increase in the third trial as the person receives the stimulus.

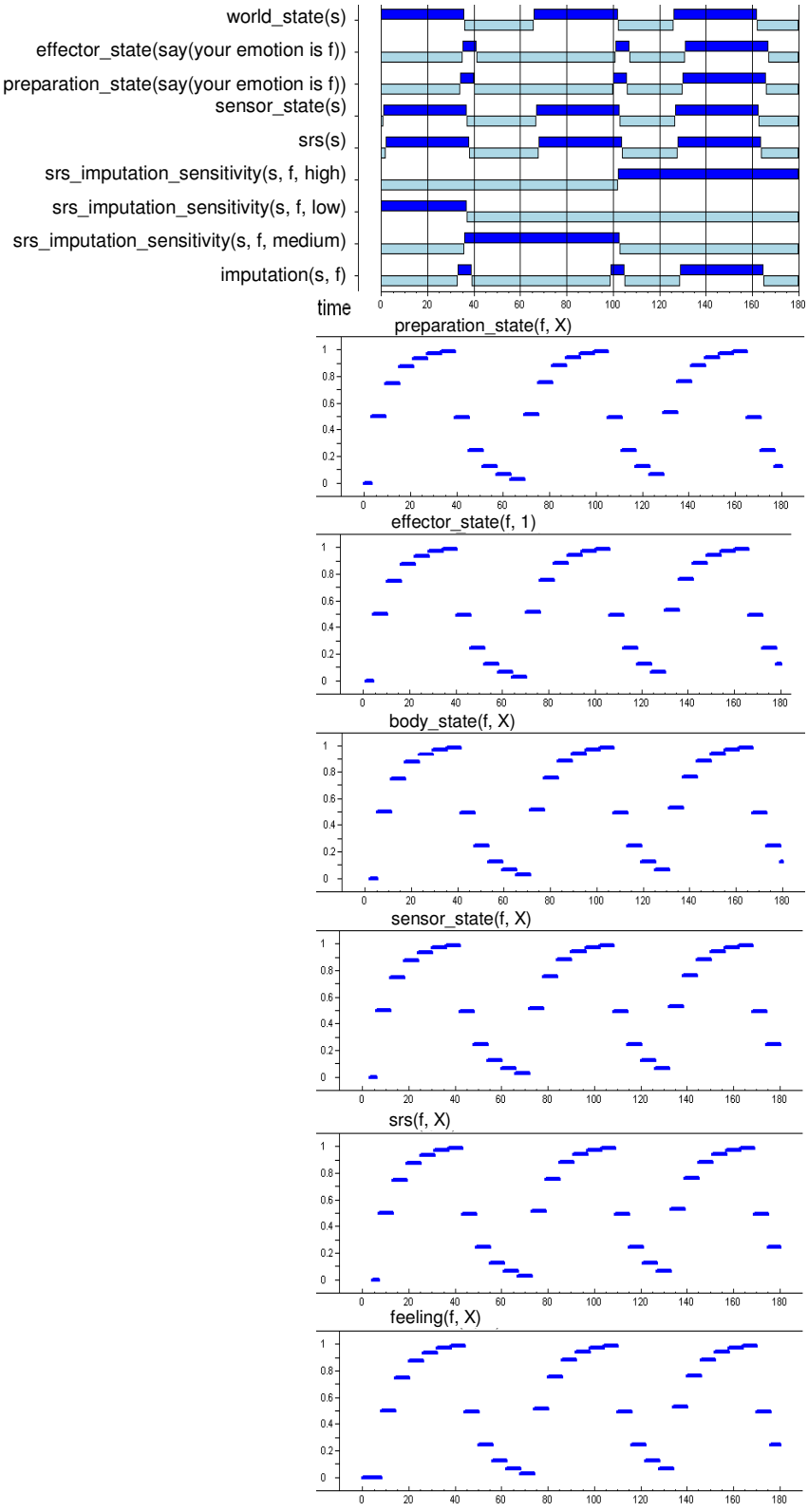

Figure 8: Simulation results for the cognitive model for adaptive emotion reading

7. Verification of Properties

To verify whether the overall behaviour of the model is according to expectations, some hypotheses (in terms of logical dynamic properties) have been identified, formally specified, and verified for simulation traces. These properties express proper emotion reading, and some of them are meant to distinguish emotion reading in a situation before adaptation and after adaptation. In particular, before an accomplished adaptation process, upon occurrence of a stimulus, first the emotion has to be felt before the emotion reading takes place. After an adaptation process, the emotion reading takes place before the emotion is felt and therefore it will take place faster.

The modelling approach for temporal expressions is based on the Temporal Trace Language TTL for formal specification and verification of dynamic properties; cf. Bosse, Jonker, Meij, Sharpanskykh, and Treur, 2009). This reified temporal predicate logical language supports formal specification and analysis of dynamic properties, covering both qualitative and quantitative aspects. TTL is built on atoms referring to states, time points and traces. A *state* of a process for (state) ontology Ont is an assignment of truth values to the set of ground atoms in the ontology. The set of all possible states for ontology Ont is denoted by STATES(Ont). To describe sequences of states, a fixed *time frame* T is assumed which is linearly ordered. A *trace* γ over state ontology Ont and time frame T is a mapping $\gamma : T \rightarrow$ STATES(Ont), i.e., a sequence of states γ_t ($t \in T$) in STATES(Ont). The set of *dynamic properties* DYNPROP(Ont) is the set of temporal statements that can be formulated with respect to traces based on the state ontology Ont in the following manner. Given a trace γ over state ontology Ont, the state in γ at time point t is denoted by state(γ, t). These states can be related to state properties via the formally defined satisfaction relation $\models$. Then, state(γ, t) $\models$ p denotes that state property p (from sort SPROP(Ont)) holds in trace γ at time t. Based on these statements, dynamic properties can be formulated in a sorted first-order predicate logic, using quantifiers over time and traces and the usual first-order logical connectives such as $\neg$, $\wedge$, $\vee$, $\Rightarrow$, $\forall$, $\exists$. A special software environment has been developed for TTL, featuring a Property Editor for building TTL properties and a Checking Tool that enables formal verification of such properties against a set of traces.

Using the TTL environment, the following Global Properties (GP's) have been identified, formalised and automatically verified against various simulation traces (first an abbreviation is introduced to count how often a state holds in a certain time period):

Abbreviations

state_holds_times_between(S:SPROP,0,tb,te:TIME,γ:TRACE) $\equiv$ $\neg$ [$\exists$t1:TIME tb<t1<te & state(γ, t1) |= S]

state_holds_times_between(S:SPROP,n+1,tb,te:TIME,γ:TRACE) $\equiv$
$\exists$t1:TIME tb<t1<te &
state(γ, t1) |= S & $\neg$[$\exists$t2:TIME tb<t2<t1 & state(γ, t2) |= S] & state_holds_times_between(S, n, t1, te, γ)

GP1a Input-Output Correlation Timing

In trace γ, if at time point t1 the person perceives a facial expression of another person,
then within time duration D this leads to communication about the person's emotional state.

GP1a(t1:TIME, γ:TRACE, D:REAL) $\equiv$
state(γ, t1) |= sensor_state(othersface(F)) $\Rightarrow$ [$\exists$t2:TIME t1<t2<t1+D & state(γ, t2) |= effector_state(your emotion is F)]

This first property checks whether the process of responding (verbally) to the stimulus is performed correctly. As could be expected, this property indeed turned out to hold for all

simulation traces, for any t1. As an illustration, consider the trace shown in Figure 8. For this trace, GP1a holds in the situation before learning for *D=36*, and after learning it holds already for *D=6*.

GP1b Input-Output Correlation During Learning

If in trace γ between tb and te the person perceives a facial expression of another person for n (different) time points, then within time duration D this leads to communication about the person's emotional state.

GP1b(tb, te:TIME, n:INTEGER, γ:TRACE, D:REAL) ≡

state_holds_times_between(sensor_state(othersface(F)), n, tb, te, γ) ⇒

[∃t:TIME te<t<te+D & state(γ, t) |= effector_state(your emotion is F)]

This property also holds for all traces and time points. For the trace shown in Figure 8, it holds for n=3 and D=6. Hence, in all situations that the person perceived the stimulus three times, this resulted in a response within 6 time points.

GP2 Successful Associative Learning

If in trace γ between tb and te state property S1 and S2 hold together for n (different) time points,
then eventually a relation between these states will be learned.

GP2(tb, te:TIME, n:INTEGER, γ:TRACE) ≡

∀S1,S2:SPROP

state_holds_times_between(S1∧S2, n, tb, te, γ) ⇒

[∃t:TIME ∃w:REAL te<t<te+D & state(γ, t) |= sensitivity_for_relation_between(S1,S2, w) & w>δ]

This property holds for all traces for *n=2* (and for *D=1*), which confirms that the associative learning is directly successful after two trials. Note that here δ is a certain sensitivity threshold, which can be considered to depend on n. Thus, an example instance of

sensitivity_for_relation_between(S1, S2, w)

could be the state property

srs_imputation_sensitivity(s, f, high).

GP3a Emotion reading with the person's own feeling

In trace γ, if at time point t1 a stimulus occurs,
then there is a point in time that the emotion is recognised whereas it is felt as well.

GP3a(t1:TIME, γ:TRACE) ≡

state(γ, t1) |= sensor_state(othersface(F)) ⇒

∃t2:TIME, V:REAL [t1<t2<t1+D & V>th & state(γ, t2) |= effector_state(your emotion is F) & state(γ, t2) |= feeling(F, V)]

GP3b Emotion reading without the person's own feeling

In trace γ, if at time point t1 a stimulus occurs,
then there is a point in time that the emotion is recognised whereas it is not felt (yet).

GP3b(t1:TIME, γ:TRACE) ≡

state(γ, t1) |= sensor_state(othersface(F)) ⇒ ∃t2:TIME, V:REAL [t1<t2<t1+D &

V≤0.1 & state(γ, t2) |= effector_state(your emotion is F) & state(γ, t2) |= feeling(F, V)]

These properties have been used to distinguish the phase when the person performs emotion reading with an experienced emotion from the phase without an experienced emotion. For example, for the trace depicted in Figure 8, checks pointed out that the second phase is entered at time point 126.

To conclude, although not proven exhaustively, the above checks have pointed out that the presented models satisfy a number of relevant expected properties. In addition, they allow the modeller to fine-tune the precise temporal aspects of the simulated emotion reading process.

8. An Interpretation Mapping from the Cognitive to the Neural Model

The cognitive model described in Section 5 by abstracting the following neurological principles to the cognitive level: (1) a *recursive body loop* for cognitive preparation states and feeling states, (2) the *mirrorring function* of cognitive preparation states as inspired by the notion of mirror neurons and by the Simulation Theory perspective on mindreading, and (3) a cognitive-level *Hebbian learning principle* for adaptivity in the cognitive model. Such an abstraction allows a modeller to exploit neurological knowledge to enrich models at the cognitive level, and not to work with neural models of the type as described in Section 3. The result is that now two models are available describing the same process, at the neural level, resp. cognitive level. These two models are formally defined objects, so as they are assumed to describe the same reality, a natural question is in how far they can be formally related to each other (see Figure 9).

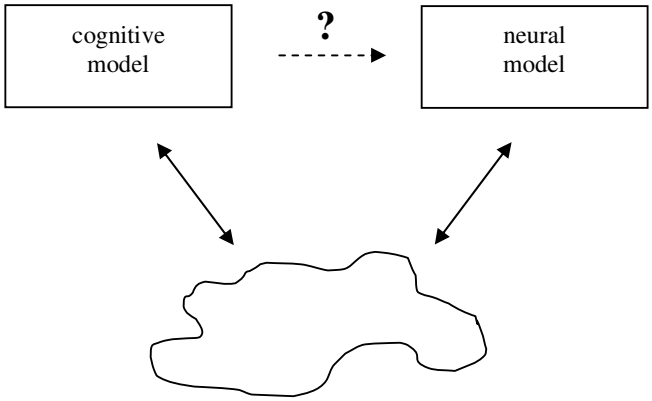

Figure 9: Relating two models describing the same reality

For cognitive models in general it is an interesting challenge to find out how they can be related to a neural and/or biological realisation. Work on this area of reduction can be found in a wide variety of publications in the philosophical literature; see, for example, (Kim, 2005). A specific reduction approach provides a particular *reduction relation*: a way in which each cognitive property *a* can be related to a neural property *b*; this *b* is often called a *realiser* for *a*. Reduction approaches differ in how these relations are defined. In (Treur, 2010) three well-known approaches are described and compared to each other: the bridge law approach, the interpretation mapping approach and the functional reduction approach, and it is shown how they can be translated into each other, when the context of the realisation is made explicit.

The notion to define reduction relations used below is the *interpretation mapping approach*; e.g., (Schoenfield, 1967, pp. 61-65). This is based on a mapping φ relating cognitive concepts *a* to neural concepts *b*, in the sense that $b = \varphi(a)$. Such a mapping is an interpretation mapping when it satisfies the property that if *L* is a cognitive law, then the statement $\varphi(L)$ can be derived from neural laws. Usually the mapping is assumed compositional with respect to connectives, for example:

$\varphi(A_1 \,\&\, A_2) = \varphi(A_1) \,\&\, \varphi(A_2)$
$\varphi(A_1 \vee A_2) = \varphi(A_1) \vee \varphi(A_2)$
$\varphi(\neg A_1) = \neg \varphi(A_1)$
$\varphi(A_1 \rightarrow A_2) = \varphi(A_1) \rightarrow \varphi(A_2)$

In this section it is shown how the cognitive model for adaptive emotion reading, has been mapped onto the neural model, by an interpretation mapping.

In order to define an interpretation mapping from the cognitive model to the neural model for adaptive emotion generation, one needs to formally settle, for example, which neural states exactly are to be interpreted as feeling the emotion, and which as the imputation of the emotion to a person. For the state properties of the cognitive model, the interpretation mapping π (indicated by a question mark in Figure 9) has been defined as follows, where a criterion for considering RN(S, F) as imputation is defined by a threshold of *0.75*.

π(S) = activated(S, 1)
π(sensor_state(S)) = activated(sensor_state(S), 1)
π(srs(S)) = activated(SN(S), 1)
π(preparation_state(F, V)) = activated(PN(F), V)
π(feeling(F, V)) = activated(FN(F), V)
π(effector_state(F, V)) = effector_state(F, V)
π(sensor_state(F, V)) = sensor_state(F, V)
π(srs(F, V)) = activated(SN(F), V)
π(imputation(S, F)) = $\exists$V V≥0.75 & activated(RN(S, F), V)
π(srs_imputation_sensitivity(S, B, V)) = $\exists$W qualifies_as(W, V) & connectedto(FN(f), SN(s), PN(f), α, W)

Here qualifies_as(W, V) is a predicate that is assumed to relate the values V used in the cognitive model to values W between *0* and *1* for the connection strength in the neural model. An example instantiation of this predicate is qualifies_as(0.95, high).

The mapping is extended to more complex (temporal) expressions in a compositional manner as follows:

$\pi(A_1 \& A_2) = \pi(A_1) \& \pi(A_2)$
$\pi(A_1 \rightarrow A_2) = \pi(A_1) \rightarrow \pi(A_2)$

Using this, the mapping maps the cognitive temporal relationships (depicted in Figure 7) between the different state properties specified in the cognitive model to neural relationships between state properties entailed by the neural model (depicted in Figure 1). For example, if *L* is the relationship

srs(s) & feeling(f, V) → preparation_state(f, (1+V)/2)

which holds in the cognitive model, then *L* is mapped by φ_1 onto

π(*L*) = π(srs(s) & feeling (f, V) → preparation_state(f, (1+V)/2))
= π(srs(s) & feeling(f, V)) → π(preparation_state(f, (1+V)/2)))
= π(srs(s)) & π(feeling(f, V)) → π(preparation_state(f, (1+V)/2)))
= activated(SN(s), 1) & activated(FN(f), V) → activated(PN(f), (1+V)/2)

The latter expression is not literally part of the neural model, but is entailed by it, in particular by

connectedto(FN(f), SN(s), PN(f), α, β)

for α=β=*0.5* together with the general rule

connectedto(X1, X2, Y, α, β) & activated(X1, V1) & activated(X2, V2) → activated(Y, α* V1+β *V2)

that specifies propagation of activation through connections. In a similar way a property has been mapped that expresses that always an emotion is imputed to a sensed stimulus: the temporal relation *L'* given by

srs(s) → imputation(s, f)

is entailed by the temporal relations in the neural model. It is mapped as follows:

$\boldsymbol{\pi}(L')$ = $\boldsymbol{\pi}$(srs(s) $\rightarrow$ imputation(s, f))
= $\boldsymbol{\pi}$(srs(s)) $\rightarrow$ $\boldsymbol{\pi}$(imputation(s, f))
= activated(SN(s), 1) $\rightarrow$ $\exists$V V≥0.75 & activated(RN(s, f), V)

Indeed this property is entailed by a connection

connectedto(FN(B), SN(S), RN(S, B), α, W)

but only when $W \geq 0.75$ (which makes the condition FN(B) superfluous to pass the threshold) and the temporal relationship

connectedto(X1, X2, Y, α, β) & activated(X1, V1) & activated(X2, V2) $\rightarrow$ activated(Y, α*V1+β*V2)

in the neural model. Note that this does not hold when W is too low, for example, when $W = 0.5$. An interpretation mapping for the communication extensions of the emotion reading model has been defined as a specialisation of the mapping φ_1 above as follows:

$\boldsymbol{\pi}$(preparation_state(say(your_emotion_is(f)))) = $\exists$V V≥0.75 & activated(PN(say(your_emotion_is(f))), V)
$\boldsymbol{\pi}$(effector_state(say(your_emotion_is(f)))) = $\exists$V V≥0.75 & activated(effector_state(say(your_emotion_is(f))), V)

The learning rule of the cognitive model has been mapped as follows:

$\boldsymbol{\pi}$(srs(S) & imputation(S, B) & srs_imputation_sensitivity(S, B, V1) & next_value(V1, V2)
$\rightarrow$ srs_imputation_sensitivity(S, B, V2))
= $\boldsymbol{\pi}$(srs(S) & imputation(S, B) & srs_imputation_sensitivity(S, B, V1) & next_value(V1, V2))
$\rightarrow$ $\boldsymbol{\pi}$(srs_imputation_sensitivity(S, B, V2))
= $\boldsymbol{\pi}$(srs(S)) & π(imputation(S, B)) & $\boldsymbol{\pi}$(srs_imputation_sensitivity(S, B, V1)) & $\boldsymbol{\pi}$(next_value(V1, V2)))
$\rightarrow$ $\boldsymbol{\pi}$(srs_imputation_sensitivity(S, B, V2))
= activated(SN(S), 1) & $\exists$V V≥0.75 & activated(RN(S, F), V) &
$\exists$W qualifies_as(W, V1) & connectedto(FN(f), SN(s), PN(f), α, W) & next_value(V1, V2)
$\rightarrow$ $\exists$W qualifies_as(W, V2) & connectedto(FN(f), SN(s), PN(f), α, W)

In principle this is entailed by the Hebbian learning rule

connectedto(X1, X2, Y, W, β) & activated(X1, V1) & activated(Y, V2)
$\rightarrow$ connectedto(X1, X2, Y, W+ηV1V2(1-W) - ζW, β)

in the neural model, but this also depends on the precise definition of the values in the cognitive model and the 'next value' relation. One case in which it holds is when the values for the cognitive model are exactly the same as in the neural model.

9. Discussion

In the literature on emotion reading, it is often assumed that a person uses observations of another person's body (for example facial expressions) as a basis for the emotion reading process. Models for emotion reading by a person can be of two types: either they make use of

the person's own emotion states , or they are independent of them. Models for emotion reading of the second type are available using a specific classification procedure. Here, for example, a specific emotion reading process can be modelled in the form of a prespecified classification process of facial expressions in terms of a set of possible emotions; see, for example, (Cohen, Garg, and Huang, 2000; Malle, Moses, and Baldwin, 2001; Pantic and Rothkrantz, 1997, 2000). Also models of an observing person based on reasoning based on models of the observed person are of the second type, for example (Bosse, Memon, and Treur, 2007a, 2007b). Such models are considered in the Theory Theory perspective on mindreading (e.g., Goldman, 2006). A model based on such a classification procedure or based on reasoning is able to perform emotion reading. However, within such an approach the imputed emotions will not have any relationship to a person's own emotions.

Instead, the Simulation Theory perspective on mindreading assumes that the person's own mental states are used to simulate the other person's corresponding mental states; (e.g., Goldman, 2006; Goldman and Sripada, 2004; Bosse, Memon, and Treur, 2008). In recent years, an increasing amount of neurological evidence is found that supports the Simulation Theory perspective on emotion reading, e.g., (Rizzolatti, Fogassi, and Gallese, 2001; Wohlschlager and Bekkering, 2002; Kohler, Keysers, Umilta, Fogassi, Gallese, and Rizzolatti, 2002; Ferrari, Gallese, Rizzolatti, and Fogassi, 2003; Rizzolatti, 2004; Rizzolatti and Craighero, 2004; Iacoboni, 2005, 2008). According to such a type of approach, in order to recognise emotions of other persons, humans exploit observations of these other persons' body states in order to mirror these states in the persons' own preparation states, and based on this simulation of the other person's states takes place making use of counterparts of these states.

The *first research question* was formulated in the introduction section in the following manner:

- How can emotion reading by a person be modelled taking his or her own emotional states into account, and how can this be integrated in an adaptive manner with emotion reading without taking into account the person's own emotional states?

This question was addressed by the models presented in the current paper integrating approaches to mindreading of the two types. The models do not discriminate between different emotions; they are based on the notions of (preparatory) mirror neurons and Damasio's perspective on emotions and feelings based on a recursive body loop (cf. Damasio, 1999, 2003), generating a converging positive feedback loop based on reciprocal causation between mirrorring preparation states and feeling states. The models were equipped with an adaptation model to learn a direct connection between sensory representation of a stimulus and emotion imputation. Thus, after a learning phase the person can perform emotion reading without taking the person's own emotions into account. As this learnt pathway bypasses the person's own emotion generation process, such a direct connection is faster (it may take place within hundreds of milliseconds) than a connection via a body loop (which usually takes seconds). This time difference implies that first the emotion is recognised without feeling the corresponding person's own emotion, but within seconds the corresponding person's own emotion is in a sense added to the recognition. When an as if body loop is used instead of a body loop, the time difference will be smaller, but may still be present. An interesting question is whether it is possible to design experiments that show this time difference as predicted by the neural agent

model. As the person's own emotions are not involved anymore, it can be argued that the learnt model for emotion reading by itself is not a model from the Simulation Theory perspective, whereas the model for the learning process to obtain this model is. It may also be considered that the learnt model (or part of it) is innate, and is only further tuned by the learning process. One step further, one could even argue that the learnt part of the model fits in the Theory Theory perspective. However, notice that what is learned is only a specific and simple form of a Theory Theory model. A further exploration of the relation between adaptive emotion reading models from a Simulation Theory perspective and Theory Theory models is left for future work.

The models have been specified in LEADSTO and in Matlab. The neural model consists of two types of general rules: one for propagation of activation levels between connected neurons, and one for strengthening of connections between neurons that are active simultaneously. These rules are applied to all nodes in the network. To perform a particular simulation, only the initial activation levels and connection strengths have to be specified. The simulations performed indicated that the models are indeed able to simulate various patterns of adaptive emotion reading. An interesting challenge for the future is to extend the models such that they can cope with multiple qualitatively different emotional stimuli (e.g., related to joy, anger, or fear), and their interaction.

Some other computational models related to mirror neurons are available in literature; for instance: a genetic algorithm model which develops networks for imitation while yielding mirror neurons as a byproduct of the evolutionary process (Borenstein and Ruppin, 2005); the mirror neuron system (MNS) model that can learn to 'mirror' via self-observation of grasp actions (Oztop and Arbib, 2002); the mental state inference (MSI) model that builds on the forward model hypothesis of mirror neurons (Oztop, Wolpert, and Kawato, 2005). A comprehensive review of these computational studies can be found in (Oztop, Kawato, and Arbib, 2006). All of the above listed computational models (and many others available in the literature) are targeted to imitation, whereas the neural model presented here specifically targets to interpret somebody else's emotions.

The *second research question* in the introduction section was formulated in the following manner:

- How can state of the art neurological knowledge be exploited in modelling these emotion reading processes; how can they be modelled at a neural level and how in a more abstracted form at a cognitive level, and how do the obtained models at these two levels relate to each other?

This question was addressed by providing both a neural level model and a cognitive level model, illustrating the possibilities. Modeling causal relations discussed in neurological literature in a cognitive level model does not take specific neurons into consideration but can use more abstract mental states. This is a way to use results from the large and more and more growing amount of neurological literature, without abandoning the cognitive modelling level. This method can be considered as lifting neurological knowledge to a cognitive level of description. In a more detailed manner, Bickle (1998, pp. 205-208), illustrates a similar perspective for the higher level (e.g., folk psychological) in relation to the lower-level (e.g., neurobiological) explanation in the context of Hawkin and Kandel's (1984a,1984b) case; see also (Jonker, Treur, and Wijngaards, 2002):

> 'The abstract processing structure of the two networks is very similar, at least at a coarse-grained level of analysis. The gross causal flow, from sensations through representational states to behavior, is mostly the same. Imagine the two accounts diagrammed as a set of nodes, with each node representing a representational state occurring in the explanation, connected by arrows representing the causal effects. If we overlay the nets, landmark nodes and arrows of the two would largely lie one on top of the other. (…) Of course, the functional profiles assigned to cognitive states on Hawkin and Kandel's neurobiological account are much more fine-grained and detailed, for that account recognizes distinctions and connections that folk psychology either lumps together or leaves extremely vague (…) Here again, however, we can expect that injection of some neurobiological details back into folk psychology would fruitfully enrich the latter, and thus allow development of a more fine-grained folk-psychological account that better matches the detailed functional profiles that neurobiology assigns to its representational states. There is no principled reason against such enrichment.' (Bickle, 1998, p. 207-208)

Here Bickle suggests that by relating a folk psychological explanation to a neurobiological account, a decision can be made to enrich the former, based on the more detailed account provided by the latter. Note that what he sketches about to 'overlay the nets' visualises quite well the interpretation mapping defined in Section 8, which can be visualised as a mapping from the cognitive 'net' depicted in Figure 7 to the neural 'net' in Figure 1.

The type of cognitive level model that results from adopting principles from the neurological level may inherit some characteristics (in the technical and/or conceptual sense) from the neurological level. For example, it takes *cognitive states as having a certain activation level*, instead of binary (to occur or not to occur). This is needed to be able to model *gradual adaptation processes* and *loops*, which both are essential for the processes addresssed here, but are not always covered by (symbolic) cognitive modelling approaches. As a consequence, for a cognitive state depending on multiple other states, values for such activation levels have to be combined, to obtain an activation level for this state. Therefore combination functions are needed, for example, as a technique to determine the level of the preparation state from the levels of sensory representations of the stimulus and of the body. However, the technique used for modelling is not to be considered a distinguishing criterion between neural or cognitive modelling level. In order to incorporate at the cognitive level elements put forward by neuroscience, such as gradual adaptation and loops, modelling techniques at the cognitive level are needed that maybe usually are associated to neural modelling practice; but techniques themselves are neutral in for what they are used, be it at a cognitive or at a neural level.

Another example is the notion of mirror neurons, discovered in neurological context. The function of mirroring can be abstracted to a comparable function of a state at the cognitive level as shown here: a *mirroring function of a cognitive preparation state*. Yet another example is the Hebbian learning principle, which originally was formulated for neurons, but can easily be abstracted to a *cognitive Hebbian learning principle*, as was done here. So, in order to model an adaptive agent at a cognitive level abstracting from neurological detail, still some machinery may be needed that may usually be associated to a neural modelling perspective. In order to obtain cognitive models with more complex, adaptive and human-like behaviour, the toolset for the modeller has to include such numerical modelling techniques, enabling to model in a hybrid logical/numerical manner.

To show how the more abstract adaptive cognitive model for emotion reading is related to the neurological context, it was formally related to the neural model. This adaptive neural model makes use of mirror neurons, and learns a direct (synaptic) connection between sensory neurons (for example, concerning another person's face expression) and the emotion recognition

neurons. Based on the literature on reduction such as (Kim, 2005; Treur, 2010), it was shown how the models can be related to each other by an interpretation mapping. This interpretation mapping was first defined on state properties, and then extended by compositionality to dynamic relations. For some of the state properties it was needed that a qualitative variable at the cognitive level was related to a quantitative variable at the neurological level, by using as a condition that the value of the quantitative variable was above some threshold. At these points it the difference in abstraction level between the two models is shown.

A third research question that can be formulated concerns the assessment of neurological theories and their relations to empirical data, and what is the role of computational models such as the ones presented here:

- How can the introduced computational models play a role in strengthening the assessment and validation of neurological theories?

Although usually inspired by empirical results, scientific theories always have a certain extent of being speculative. It is interesting and useful to make further analyses and assessments of theories that are active in the research community, as the current paper does. Indeed Damasio (1999, 2003)'s theories and the theories about mirror neuron systems (Rizzolatti and Sinigaglia, 2008; Pineda, 2009; Iacoboni, 2008) used as a basis here currently are active; they occur in current (cognitive) neuroscience textbooks (e.g., Gazzaniga, 2009; Purves et al., 2008; Ward, 2010), and in many other state of the art publications. It is a joint effort of a whole multidisciplinary research community to assess such theories both by formal analysis methods and by empirical research. Different research groups play different roles in such a process, taking into account their specific background and expertise; some may take more responsibility for contributing empirical research, some other more for contributing computational modelling approaches and formal analyses. The joint effort has as an aim over a longer time period to bring all of these aspects further. Although the authors fully recognize the third research question formulated above as being important in all of its aspects, the current paper focuses on a contribution from the latter side.

Computational modelling techniques play a useful tool role for analysis of theories, as they can be used to determine in a precise manner the implications of a theory. For example, by simulation or formal verification it can be determined in a detailed manner which patterns may or may not emerge from basic mechanisms described by such a theory. This paper has indicated how the idea of a recursive body loop can be integrated with the notion of mirror neurons and Hebbian learning, with resulting patterns that are quite plausible according to the neurological literature. In this sense the models contribute a positive evaluation of these theories. This is also a relative validation of the models themselves, with respect to the neurological literature. Validation of the theories based on precise empirical data by using the presented models is an interesting challenge, and not impossible, but also not trivial. Such a more extensive empirical evaluation of the theories and models as presented is left for future work.

The role of more extensive cognitive interpretation or labeling as an ingredient for a specific emotion has not been taken in the scope of this model, as the current aim was to follow Damasio's theory. However, an interesting extension to be addressed in future work would be to incorporate such cognitive interpretation as an extension in the models.

References

Baron-Cohen, S. (1995). *Mindblindness*. MIT Press

Bi, G.Q., and, Poo, M.M. (2001) Synaptic Modifications by Correlated Activity: Hebb's Postulate Revisited. Ann Rev Neurosci, vol. 24, pp. 139-166.

Bickle, J. (1998). *Psychoneural Reduction: The New Wave*. MIT Press, Cambridge, Massachusetts.

Bogdan, R.J. (1997). *Interpreting Minds*. MIT Press

Borenstein, E., & Ruppin, E. (2005). The evolution of imitation and mirror neurons in adaptive agents. *Cognitive Systems Research,* 6(3), pp. 229-242.

Bosse, T., Jonker, C.M., Meij, L. van der, Sharpanskykh, A, and Treur, J. (2009). Specification and Verification of Dynamics in Cognitive Agent Models. *International Journal of Cooperative Information Systems*, vol. 18, 2009, pp. 167 - 193.

Bosse, T., Jonker, C.M., Meij, L. van der, and Treur, J. (2007). A Language and Environment for Analysis of Dynamics by Simulation. *International Journal of Artificial Intelligence Tools*, vol. 16, 2007, pp. 435-464.

Bosse, T., Jonker, C.M., and Treur, J., (2008). Formalisation of Damasio's Theory of Emotion, Feeling and Core Consciousness. *Consciousness and Cognition Journal*, vol. 17, 2008, pp. 94–113.

Bosse, T., Memon, Z.A., and Treur, J., (2007a). A Two-level BDI-Agent Model for Theory of Mind and its Use in Social Manipulation. In: P. Olivier, C. Kray (eds.), *Proceedings of the Artificial and Ambient Intelligence Conference, AISB'07, Mindful Environments Track*. AISB Publications, 2007, pp. 335-342.

Bosse, T., Memon, Z.A., and Treur, J., (2007b). Emergent Storylines Based on Autonomous Characters with Mindreading Capabilities. In: Lin, T.Y., Bradshaw, J.M., Klusch, M., Zhang, C., Broder, A., and Ho, H.(eds.), *Proceedings of the Seventh IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT'07*. IEEE Computer Society Press, 2007, pp. 207-214.

Bosse, T., Memon, Z.A., and Treur, J. (2008), Adaptive Estimation of Emotion Generation for an Ambient Agent Model. In: Aarts, E., Crowley, J.L., Ruyter, B. de, Gerhäuser, H., Pflaum, A., Schmidt, J., Wichert, R. (eds.), *Ambient Intelligence, Proceedings of the Second European Conference on Ambient Intelligence, AmI'08*. Lecture Notes in Computer Science, vol. 5355. Springer Verlag, 2008, pp. 141–156.

Bosse, T., Memon, Z.A., and Treur, J. (2009a), A Neural Model for Adaptive Emotion Reading Based on Mirror Neurons and Hebbian Learning, In: *9th International Conference on Cognitive Modelling,* ICCM 2009.

Bosse, T., Memon, Z.A., and Treur, J. (2009b), An Adaptive Emotion Reading Model. In: Taatgen & H. van Rijn (eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, (pp. 1006-1011). Austin, TX: Cognitive Science Society CogSci'09.

Bosse, T., Memon, Z.A., and Treur, J. (2009c), A Adaptive Agent Model for Emotion Reading by Mirroring Body States and Hebbian Learning. In: Yang, J.-J.; Yokoo, M.; Ito, T.; Jin, Z.; Scerri, P. (eds.). *Proceedings of the 12th International Conference on Principles of Practice in Multi-Agent Systems, PRIMA'09.* Lecture Notes in Artificial Intelligence, vol. 5925. Springer Verlag, 2009, pp. 552-562.

Carr, L., Iacoboni, M., Dubeau, M.C., Mazziotta, J.C., Lenzi, G.L. (2003). Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas. Proc Natl Acad Sci USA, vol. 100, pp. 5497-5502.

Cohen, I., Garg, A., and Huang, T.S. (2000). Emotion recognition using multilevel HMM. In: *Proceedings of the NIPS Workshop on Affective Computing*, Colorado, 2000.

Dapretto, M., Davies, M.S., Pfeifer, J.H., Scott, A.A., Sigman, M., Bookheimer, S.Y., and Iacoboni, M. (2006). Understanding emotions in others: Mirror neuron dysfunction in children with autism spectrum disorder. Nature Neuroscience, vol. 9, pp. 28-30.

Damasio, A. (1999). *The Feeling of What Happens: Body, Emotion and the Making of Consciousness*. Harcourt Brace, 1999.

Damasio, A. (2003). *Looking for Spinoza*: *Joy, Sorrow, and the Feeling Brain.* Vintage books, London, 2003.

Dennett, D.C. (1987). *The Intentional Stance*. MIT Press. Cambridge Mass.

Ferrari PF, Gallese V, Rizzolatti G, Fogassi, L. 2003. Mirror neurons responding to the observation of ingestive and communicative mouth actions in the monkey ventral premotor cortex. *Eur. J. Neurosci.* 17:1703–14.

Gazzaniga, M.S. (ed.) (2009). *The Cognitive Neurosciences IV*. MIT Press.

Gerstner, W., and Kistler, W.M. (2002). Mathematical formulations of Hebbian learning. Biol. Cybern., vol. 87, 2002, pp. 404-415.

Gleitman, H. (1999). *Psychology*. W.W. Norton & Company, New York, 1999.

Goldman, A.I. (2006). *Simulating Minds: the Philosophy, Psychology and Neuroscience of Mindreading*. Oxford University Press.

Goldman, A.I., and Sripada, C.S. (2004). Simulationist models of face-based emotion recognition. *Cognition*, vol. 94, pp. 193–213.

Hawkins, R.D., and Kandel, E.R. (1984a). Is There a Cell-Biological Alphabet for Simple Forms of Learning? *Psychological Review*, vol. 91, pp. 375-391.

Hawkins, R.D., and Kandel, E.R. (1984b). Steps Toward a Cell-Biological Alphabet for Elementary Forms of Learning. In: G. Lynch, J.L. McGaugh, and N.M. Weinberger (eds.), *Neurobiology of Learning and Memory*, Guilford Press, New York, pp. 385-404, vol. 91, pp. 375-391

Hebb, D.O. (1949). *The Organization of Behaviour*. John Wiley & Sons, New York, 1949.

Iacoboni M. (2008). *Mirroring People: the New Science of How We Connect with Others*. New York: Farrar, Straus & Giroux

Iacoboni, M., (2005). Understanding others: imitation, language, empathy. In: Hurley, S. & Chater, N. (2005). (eds.). *Perspectives on imitation: from cognitive neuroscience to social science* vol. 1, MIT Press, pp. 77-100.

Jonker, C.M., Treur, J., and Wijngaards, W.C.A., (2002). Reductionist and Antireductionist Perspectives on Dynamics. *Philosophical Psychology Journal*, vol. 15, 2002, pp. 381-409.

Kim, J. (2005). *Physicalism, or Something Near Enough.* Princeton University Press, Princeton.

Kohler, E., Keysers, C., Umilta, M.A., Fogassi, L., Gallese, V., and Rizzolatti, G. 2002. Hearing sounds, understanding actions: action representation in mirror neurons. *Science* 297:846–48

Malle, B.F., Moses, L.J., Baldwin, D.A. (2001). *Intentions and Intentionality: Foundations of Social Cognition.* MIT Press.

Memon, Z.A., and Treur, J. (2008), Cognitive and Biological Agent Models for Emotion Reading. In: Jain, L., Gini, M., Faltings, B.B., Terano, T., Zhang, C., Cercone, N., Cao, L. (eds.), *Proceedings of the 8th IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT'08*. IEEE Computer Society Press, 2008, pp. 308-313.

Oztop, E., & Arbib, M.A. (2002). Schema design and implementation of the grasp-related mirror neuron system. *Biological Cybernetics,* 87(2), 116-140.

Oztop, E., Kawato, M. and Arbib M. (2006), Mirror neurons and imitation: a computationally guided review, Neural Networks 19 (3) (2006) 254-271.

Oztop, E., Wolpert, D., & Kawato, M. (2005). Mental state inference using visual control parameters. C*ognitive Brain Research,* 22(2), 129-151.

Pantic, M., and Rothkrantz, L.J.M. (1997). Automatic Recognition of Facial Expressions and Human Emotions. In: *Proceedings of ASCI'97 conference*, ASCI, Delft, pp. 196-202.

Pantic, M., and Rothkrantz, L.J.M. (2000), Expert System for Automatic Analysis of Facial Expressions, *Image and Vision Computing Journal*, vol. 18, pp. 881-905.

Pineda, J.A. (ed.), (2009). *Mirror Neuron Systems: the Role of Mirroring Processes in Social Cognition.* Humana Press Inc.

Port, R.F., Gelder, T. van (eds.), (1995). *Mind as Motion: Explorations in the Dynamics of Cognition.* MIT Press, Cambridge, Mass, 1995.

Purves, D., et al. (2008). *Principles of Cognitive Neuroscience*. Sinauer Associates.

Rizzolatti, G. and Craighero, L. (2004) The mirror-neuron system. *Annu. Rev. Neurosci.* **27**, 169–92.

Rizzolatti G, Fogassi L, Gallese V (2001). Neuro-physiological mechanisms underlying the understanding and imitation of action. Nature Rev Neurosci 2:661–670.

Rizzolatti G. 2005. The mirror-neuron system and imitation. In: Hurley, S. & Chater, N. (2005). (eds.). *Perspectives on imitation: from cognitive neuroscience to social science*, vol. 1, MIT Press, pp. 55-76.

Rizzolatti, G, and Sinigaglia, C., (2008). *Mirrors in the Brain: How Our Minds Share Actions and Emotions*. Oxford Univsersity Press, 2008.

Treur, J., (2009). On the Use of Reduction Relations to Relate Different Types of Agent Models. *Web Intelligence and Agent Systems Journal*, 2010, to appear.

Ward, J. (2010). The Student's Guide to Cognitive Neuroscience, 2nd ed. Psychology Press.

Wasserman, P.D. (1989). *Neural Computing: Theory and Practice*. Van Nostrand Reinhold, New York, 1989.

Wohlschlager A, Bekkering H. 2002. Is human imitation based on a mirror-neurone system? Some behavioural evidence. *Exp. Brain Res.* 143:335–41.

Appendix A Equilibrium Analysis for the Neural Model

Equilibrium equations for the non-adaptive case

The neural model description in the form of a system of differential equations has been used for an analysis of equilibria that can occur. Here the external stimulus level for s is assumed constant. Moreover, it is assumed that $\gamma > 0$. In general putting $\Delta y_i(t) = 0$ provides the following set of equations for $i \in N'$:

$$y_i = g(\Sigma_{j \in N} \; w_{ji} \, y_j)$$

For the given network structure these equilibrium equations are:

$$y_1 = g(w_{01} \, y_0) \qquad y_2 = g(w_{12} \, y_1) \qquad y_4 = g(w_{34} \, y_3) \qquad y_5 = g(w_{45} \, y_4) \qquad y_6 = g(w_{56} \, y_5)$$

$$y_7 = g(w_{67} \, y_6) \qquad y_8 = g(w_{78} \, y_7) \qquad y_3 = g(w_{23} \, y_2 + w_{83} \, y_8) \qquad y_9 = g(w_{29} \, y_2 + w_{89} \, y_8)$$

Taking into account that connections between nodes among which at least one is not a neuron have weight 1 and g the identity function, it follows that the equilibrium equations are:

$$y_2 = y_1 = y_0 \qquad y_7 = y_6 = y_5 = y_4 = y_3 \qquad y_8 = g(w_{78} \, y_7)$$

$$y_3 = g(w_{23} \, y_2 + w_{83} \, y_8) \qquad y_9 = g(w_{29} \, y_2 + w_{89} \, y_8)$$

For the values taken in the simulation in Section 3.1, the equilibrium equations are:

$$y_2 = y_1 = y_0 \qquad y_7 = y_6 = y_5 = y_4 = y_3$$

$$y_8 = g(0.5 \, y_7) \qquad y_3 = g(0.1 \, y_2 + 0.1 \, y_8) \qquad y_9 = g(0.1 y_8)$$

As the threshold was taken 0.1 it follows from the equations that for stimulus level $y_0 = 0$ all values for y_i are (almost) 0, and for stimulus level $y_0 = 1$ that all values for y_i are 1, which is also shown by the simulation in Figure 4.

Equilibrium equations for the adaptive case

Also for the adaptive case, equilibrium equations have been found. Here it is assumed that $\gamma, \eta > 0$. Putting both $\Delta y_i(t) = 0$ and $\Delta w_{ij}(t) = 0$ provides the following set of equations for $i, j \in N'$:

$$y_i = g(\Sigma_{j \in N} \; w_{ji} \, y_j) \qquad\qquad y_i y_j (1 - w_{ij}) - \zeta w_{ij} = 0$$

For $\zeta = 0$ from the latter set of equations (second line), it immediately follows that for any pair $i, j \in N'$ it holds: either $y_i = 0$ or $y_j = 0$ or $w_{ij} = 1$. In particular, when for an equilibrium state both y_i and y_j are nonzero, then $w_{ij} = 1$. In simulations such as the one shown in Section 3.2, when a constant stimulus level 1 is taken, an equilibrium state is reached in which learned connection strength $w_{ij} = 1$, and all y_i are 1. For the general case with $\zeta \neq 0$ in an equilibrium state it holds:

$$w_{ij} = \frac{\eta y_i y_j}{\eta y_i y_j + \zeta}$$

When $y_i \, , \, y_j \neq 0$, the above equation is equivalent to:

$$w_{ij} = \frac{1}{1 + \frac{\zeta}{\eta y_i y_j}}$$

From this it follows that:

$$w_{ij} \leq \frac{1}{1 + \frac{\zeta}{\eta}} < 1$$

In the simulation examples with nonzero extinction rate, this upper bound indeed can be observed.

Appendix B Specification of the Adaptive Cognitive Model

LP1 Sensing a stimulus

If stimulus S occurs then a sensor state for S will occur.

world_state(S) ↠ sensor_state(S)

LP2 Generating a sensory representation of a stimulus

If a sensor state for S occurs, then a sensory representation for S will occur.

sensor_state(S) ↠ srs(S)

LP3 From sensory representation and emotion to preparation*

If a sensory representation for *s* occurs and feeling *b* has level V,

then the preparation state for body state *b* will occur with level (1+V)/2.

srs(s) & feeling(b, V) ↠ preparation_state(b, (1+V)/2)

If no sensory representation for *s* occurs and feeling *b* has level V,

then preparation state for body state *b* will occur with level V/2.

not srs(s) & feeling(b, V) ↠ preparation_state(b, V/2)

LP4 From preparation to body modification

If preparation state for body state *B* occurs with level V,

then the body state is modified to express *B* with level V.

preparation_state(B, V) ↠ effector_state(B, V)

LP5 From body modification to modified body

If the body state is modified to express *B* with level V, then the body state will have expression *B* with level V.

effector_state(B, V) ↠ body_state(B, V)

LP6 Sensing a body state

If body state B with level V occurs, then body state is sensed.

body_state(B, V) ↠ sensor_state(B, V)

LP7 Generating a sensory representation of a body state

If body state B of level V is sensed, then a sensory representation for body state B with level V will occur.

sensor_state(B, V) ↠ srs(B, V)

LP8 From sensory representation of body state to feeling the emotion

If a sensory representation for body state B with level V occurs, then body state B is felt with level V.

srs(B V) ↠ feeling(B, V)

LP9 Imputation

If a certain body state *B* is felt, with level ≥ th and a sensory representation for *S* occurs,

then emotion *B* will imputed to *S*.

Here, th is a (constant) threshold for imputation of emotion. In the simulations shown, th is assumed 0.95.

srs(S) & feeling(B, V) & V≥th ↠ imputation(S, B)

LP4* From preparation to sensory representation of body state

If preparation state for body state B occurs with level V,

then a sensory representation for body state B with level V will occur.

preparation_state(B, V) ↠ srs(B, V)

LP10 Communication preparation

If emotion *B* is imputed to *S*, then a related communication is prepared

imputation(B, S) ↠ preparation_state(say(your emotion is B))

LP11 Communication

If a communication is prepared, then this communication will be performed.

preparation_state(say(your emotion is B)) ↠ effector_state(say(your emotion is B))

LP12 Direct imputation

If the imputation sensitivity between *S* and *B* is high
and a sensory representation for *S* occurs,
then emotion *B* will imputed to *S*.

srs(S) & srs_imputation_sensitivity(S, B, high) → imputation(S, B)

LP13 Imputation sensitivity adaptation

If the imputation sensitivity from *S* to *B* is *W1*
and a sensory representation for *S* occurs
and an imputation occurs for *B* to *S*,
then the imputation sensitivity will become the value *W2* next to *W1*.

srs(S) & imputation(S, B) & srs_imputation_sensitivity(S, B, W1) & next_value(W1, W2)
→ srs_imputation_sensitivity(S, B, W2)

LP14 Imputation sensitivity persistence

If the imputation sensitivity is *W1* and no increase occurs,
then it will remain the same.

srs_imputation_sensitivity(S, B, W1) & next_value(W1, W2) &
not srs_imputation_sensitivity(S, B, W2) → srs_imputation_sensitivity(S, B, W1)

srs_imputation_sensitivity(S, B, high) → srs_imputation_sensitivity(S, B, high)