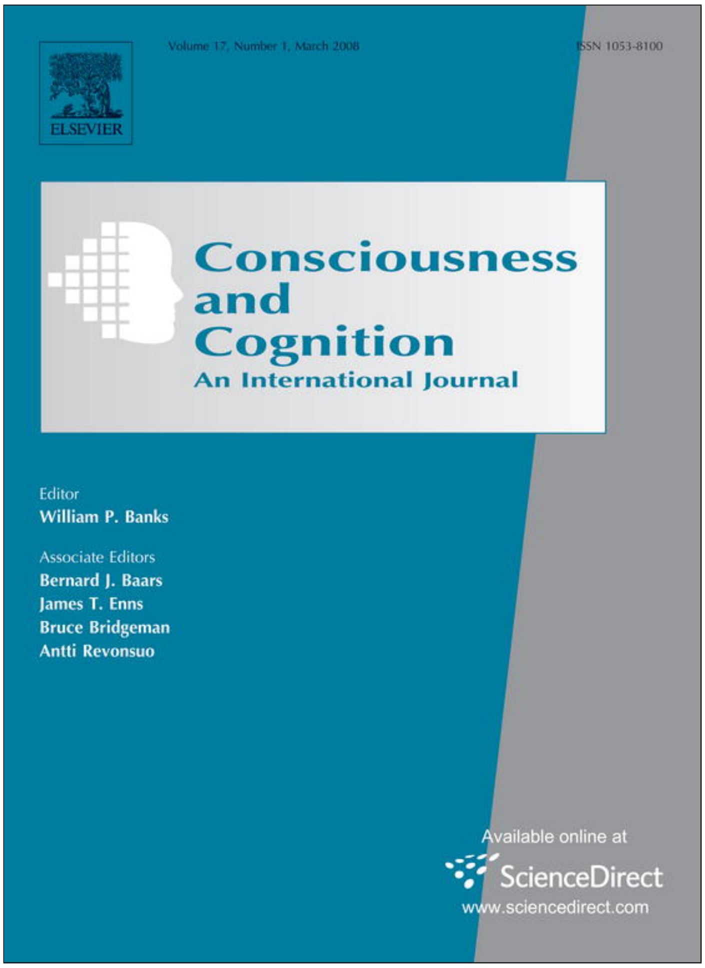
Volume 17, Number 1, March 2008
ISSN 1053-8100
ELSEVIER
Consciousness and Cognition
An International Journal
Editor
William P. Banks
Associate Editors
Bernard J. Baars
James T. Enns
Bruce Bridgeman
Antti Revonsuo
Available online at
ScienceDirect
www.sciencedirect.com

Available online at www.sciencedirect.com
ScienceDirect
Consciousness and Cognition 17 (2008) 94–113
Consciousness and Cognition
www.elsevier.com/locate/concog

Formalisation of Damasio's theory of emotion, feeling and core consciousness

Tibor Bosse [a], Catholijn M. Jonker [b], Jan Treur [a,*]

[a] Vrije Universiteit Amsterdam, Department of Artificial Intelligence, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
[b] Delft University of Technology, Department of Mediametics, Man Machine Interaction Group,
Mekelweg 4, 2628 CD Delft, The Netherlands

Received 25 August 2006
Available online 8 August 2007

Abstract

This paper contributes an analysis and formalisation of Damasio's theory on core consciousness. Three important concepts in this theory are 'emotion', 'feeling' and 'feeling a feeling' (or core consciousness). In particular, a simulation model is described of the dynamics of basic mechanisms leading via emotion and feeling to core consciousness, and dynamic properties are formally specified that hold for these dynamics at a more global level. These properties have been automatically checked for the simulation model. Moreover, a formal analysis is made of relevant notions of representation used by Damasio. As part of this analysis, specifications of representation relations have been verified and confirmed against the simulation model.

Keywords: Modelling; Feeling; Core conciousness; Damasio

1. Introduction

In Damasio (1999) the neurologist Antonio Damasio puts forward his theory of consciousness. He describes his theory in an informal manner, and supports it by evidence from neurological practice. More experimental work supporting his theory is reported in Damasio et al. (2000), Parvizi and Damasio (2001) and Parvizi, Hoesen, van Buckwalter, and Damasio (2006). Damasio's theory is described on the one hand in terms of the occurrence of certain neural states (or neural patterns), and temporal or causal relationships between them. Formalisation of these relationships requires a modelling format that is able to express direct temporal or causal dependencies. On the other hand Damasio gives interpretations of most of these neural states as representations, for example as 'sensory representation', or 'second-order representation'. This requires an analysis of what it means that a neural state is a representation for something. This paper focuses on Damasio's notions of 'emotion', 'feeling' and 'core consciousness' or 'feeling a feeling'. Damasio

[*] Corresponding author. Fax: +31 20 598 7653.
E-mail addresses: tbosse@cs.vu.nl (T. Bosse), catholijn@mmi.tudelft.nl (C.M. Jonker), treur@cs.vu.nl (J. Treur).

1053-8100/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.concog.2007.06.006

(1999) describes an *emotion as neural object* (or *internal emotional state*) as an (unconscious) neural reaction to a certain stimulus, realised by a complex ensemble of neural activations in the brain. As the neural activations involved often are preparations for (body) actions, as a consequence of an internal emotional state, the body will be modified into an *externally observable emotional state*. Next, a *feeling* is described as the (still unconscious) sensing of this body state. Finally, *core consciousness* or *feeling* a *feeling* is what emerges when the organism detects that its representation of its own body state (the *proto-self*) has been changed by the occurrence of the stimulus: it becomes (consciously) aware of the feeling. A brief summary of the main basic assumptions underlying Damasio's approach is expressed in:

> *'In summary, the complete course of events, from emotion to feeling to feeling of feeling may be partitioned along five steps (...):*

1. *Engagement of the organism by an inducer of emotion, for instance, a particular object processed visually, resulting in visual representations of the object. The object may be made conscious or not, and may be recognized or not, because neither consciousness of the object nor recognition of the object are necessary for the continuation of the cycle.*
2. *Signals consequent to the processing of the image of the object activate neural sites that are preset to respond to the particular class of inducer to which the object belongs (emotion-induction sites).*
3. *The emotion-induction sites trigger a number of responses toward the body and toward other brain sites, and unleash the full range of body and brain responses that constitute emotion.*
4. *First-order neural maps in both subcortical and cortical regions represent changes in body state, regardless of whether they were achieved via 'body loop', 'as if body loop', or combined mechanisms. Feelings emerge.*
5. *The pattern of neural activity at the emotion-induction sites is mapped in second-order neural structures. The protoself is altered because of these events. The changes in protoself are also mapped in second-order neural structures. An account of the foregoing events, depicting a relationship between the 'emotion object' (the activity at the emotion induction sites) and the proto-self is thus organized in second-order structures.*

> *This perspective on emotion, feeling, and knowing is unorthodox. First, I am suggesting that (...) 'having a feeling' is not the same as 'knowing a feeling', that reflection on feeling is yet another step up. (...) The inescapable and remarkable fact about these three phenomena – emotion, feeling, consciousness – is their body relatedness. (...) As the representations of the body grow in complexity and coordination, they come to constitute an integrated representation of the organism, a proto-self. Once that happens, it becomes possible to engender representations of the proto-self as it is affected by interactions with a given environment. It is only then that consciousness begins, only thereafter that an organism that is responding beautifully to its environment begins to discover that it is responding beautifully to its environment. But all of these processes – emotion, feeling, and consciousness – depend for their execution on representations of the organism. Their shared essence is the body.'* (Damasio, 1999, pp. 283–284).

This paper aims at formalisations and simulation models for the three notions emotion, feeling and conscious feeling or core consciousness, as distinguished by Damasio (1999). In addition, the notion of representation used by Damasio is formally analysed in the context of different approaches to representational content from the literature on philosophy of mind. It is shown that the classical causal/correlational approach to representational content (e.g., Kim, 1996, pp. 191–193), is inappropriate to describe the notion of representation for core consciousness used by Damasio, as this notion essentially involves more complex temporal relationships describing histories of the organism's interaction with the world. An alternative approach is shown to be better suited: representational content as relational specification over time and space (cf. Kim, 1996, pp. 200–202). Criteria for this approach are formalised, and it is shown that the formalisations of Damasio's notions indeed fit these criteria.

The effort to obtain a formalisation of an as yet informally described theory can only be justified if sufficiently substantial gains can be indicated. In this respect, for the effort reported here the following contributions can be recognised:

(a) assessing the theory with respect to unintended conceptual incompleteness or ambiguity, and if present, identification (and possibly removal) of such incompletenesses or ambiguities;
(b) assessing the theory with respect to internal coherency or consistency, and if present, detection (and possibly removal) of such incoherencies or inconsistencies;
(c) increasing detailedness of the theory;
(d) deriving more detailed implications of the theory;
(e) relating the theory to other theories and perspectives in the literature;
(f) the possibility to conduct pseudo-experiments (simulations) for the theory by computer support;
(g) the possibility to use computer support for verification and validation of the theory.

The contributions on these aspects have indeed turned out substantial enough to justify the effort; in the discussion the aspects (a) to (g) above will be discussed in more detail.

In this paper, first in Section 2 the modelling approach used is briefly introduced. Next, in Sections 3–5, models are presented for the processes leading to emotion, feeling and feeling a feeling (or conscious feeling), respectively, and illustrated for a simple example. Section 6 provides the results of a simulation of these models. In Section 7 it is analysed in how far the representational content of Damasio's notions can be described by two approaches from philosophy of mind. Formalisations of some of the dynamic properties of the processes leading to emotion, feeling and feeling a feeling are presented. Next, Section 8 addresses verification. It is shown that the notions for representational content developed in Section 7 indeed hold for the simulation model. The verification is performed both by automated checks and by mathematical proof. Section 9 concludes the paper with a discussion.

2. Modelling approach

To model the making of emotion, feeling and core consciousness, dynamics play an important role. Dynamics will be described as evolution of *states* over time. The notion of state as used here is characterised on the basis of an ontology Ont defining a set of state properties that do or do not hold at a certain point in time. The modelling perspective taken is not a symbolic perspective, but essentially addresses the neural processes and their dynamics as neurological processes. This implies that states are just neurological states. To successfully model such complex processes, forms of abstraction are required; for example:

- neural states or activation patterns are modelled as single state properties;
- large multi-dimensional vectors of such (distributed) state properties are composed to one single composite state property, when appropriate; e.g. (p1, p2, ...) to p and (S1, S2, ...) to S in Section 3.

To describe the dynamics of the processes mentioned above, explicit reference is made to time. Dynamic properties can be formulated that relate a state at one point in time to a state at another point in time. A simple example is the following dynamic property specification for belief creation based on observation:

> 'at any point in time t1, if the agent observes rain at t1, then there exists a point in time t2 after t1 such that at t2 the agent has internal state property s'

Here, for example, s can be viewed as a sensory representation of the rain. To express dynamic properties in a precise manner a temporal language is used in which explicit references can be made to time points and traces: the Temporal Trace Language TTL (cf. Jonker & Treur, 2002). Here a *trace or trajectory* over a state ontology Ont is a time-indexed sequence of states described in terms of state ontology Ont. The sorted predicate logic temporal trace language TTL is built on atoms referring to, e.g., traces, time and state properties. For example, 'in the internal state of agent A in trace γ at time t property s holds' is formalised by state(γ, t, internal(A)) |= s. Here |= is a predicate symbol in the language, usually used in infix notation, which is comparable to the Holds-predicate in situation calculus (cf. Reiter, 2001). Dynamic properties are expressed by temporal statements built using the usual logical connectives and quantification (for example, over traces, time and state properties). In the overview given by Galton (2003, 2006) the language TTL can be classified as a form of reified temporal logic.

To be able to perform some (pseudo)-experiments, a simpler temporal language has been used to specify simulation models in a declarative manner. This language (the LEADSTO language; cf. Bosse, Jonker, van der Meij, & Treur, 2007) enables one to model direct temporal dependencies between two state properties in successive states. This executable format is defined as follows. Let α and β be state properties of the form 'conjunction of atoms or negations of atoms', and *e*, *f*, *g*, *h*, non-negative real numbers. In the LEADSTO language the notation $\alpha \twoheadrightarrow_{e,f,g,h} \beta$, means:

> If state property α hold for a time interval with duration g, then after some delay (between e and f) state property β will hold for a time interval of length h.

For reasons of simplicity, sometimes the timing parameters *e*, *f*, *g*, *h* are left out. For a precise definition of the LEADSTO format in terms of the language TTL, see Jonker, Treur, and Wijngaards (2003). A specification of dynamic properties in LEADSTO format has as advantages that it is executable and that it can often easily be depicted graphically.

In Sections 3–6, the LEADSTO format is used to create simulation models of the processes leading to emotion, feeling and core consciousness in terms of neural processes. Given this physical-level model and its dynamic properties, a next step is to assign representational content to (some of) the relevant state properties. For nontrivial cases representational content involves histories of interaction between organism and world (Bickhard, 1993, 2000; Jonker & Treur, 2003), and this also shows up in Damasio's theory. To specify and analyse the representational content for a number of state properties of the models and the traces they generate, the more expressive TTL format is used in Section 7. Both formats are used in Section 8.

3. Emotion

First Damasio's notion of *emotion* is addressed. He explains this notion as follows:

> *'The substrate for the representation of emotions is a collection of neural dispositions in a number of brain regions (...) They exist, rather, as potential patterns of activity arising within neuron ensembles. Once these dispositions are activated, a number of consequences ensue. On the one hand, the pattern of activation represents, within the brain, a particular emotion as 'neural object'. On the other, the pattern generates explicit responses that modify both the state of the body proper and the state of other brain regions. By so doing, the responses create an emotional state and at that point, an external observer can appreciate the emotional engagement of the organism being observed* (Damasio, 1999, p. 79).

According to this description, the substrate for the representation of an emotional state is a collection of neural dispositions in the brain, which are activated as a reaction on a certain stimulus. Once this occurs, it entails modification of both the body state and the state of other brain regions. By these events, an emotional state is created which is accessible for external observation; this state may have multiple facets or dimensions.

Assume that the music you hear is so special that it leads to an emotional state in which you show some body responses on it (e.g., shivers on your back). This process is described by executable local dynamic properties taking into account internal state property `sr(music)` for activated sensory representation of hearing the music, and a vector `(p1, p2, ...)` of preparation state properties for the activation of the body responses `(S1, S2, ...)`; see Fig. 1.

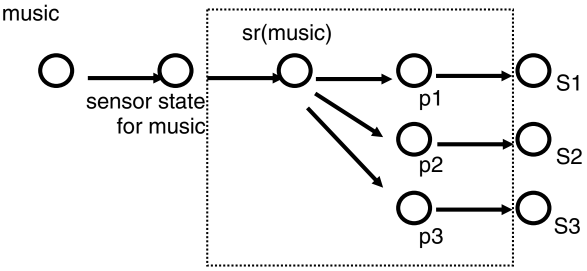

Fig. 1. Processes leading to a (multi-dimensional) emotional state.

These vectors are the possible internal emotional states. Note that the state properties are abstract in the sense that a state property refers to a specific neural activation pattern, or neural activation basin of attraction. In the model the conjunction `p1` and `p2` and .. of these preparatory state properties is denoted by `p`; this `p` can be considered a multi-dimensional composite state property. Moreover, the conjunction of the vector of all body state properties responding to the music `S1`, `S2`, ... (i.e., the respective body state properties for which `p1`, `p2`, ... are preparing) is denoted by (composite) state property `S`.

The model abstracted in this manner is depicted in Fig. 2, upper part. In formal textual format these local properties (LP's) are as follows (in the LEADSTO notation introduced in Section 2):

LP0 `music` ↠ `sensor_state(music)`
LP1 `sensor_state(music)` ↠ `sr(music)`
LP2 `sr(music)` ↠ `p`
LP3 `p` ↠ `S`

In the remainder of this paper this abstract type of modelling will be used. Notice, however, that each of the abstract state properties used are realised in the organism in a distributed manner as a large-dimensional vector of more local (neural) state properties. Also the sensory representation `sr(music)` may be considered such a composite state property with different aspects of the music represented in different forms at different places. Notice, moreover, that the names of the state properties have been chosen to support readability for humans. But in principle these names should be considered as neutral indications of neural states, such as `n1`, `n2` and so on. In Damasio (1999) the neurological mechanisms and substrates under these states and causal relationships are discussed in more detail (e.g., Damasio, 1999, pp. 47–53, pp. 59–62, pp. 79–81). For example:

> *The brain induces emotions from a remarkably small number of brain sites. Most of them are located below the cerebral cortex and are known as subcortical. The main subcortical sites are in the brain-stem region, hypothalamus, and basal forebrain. One example is the region known as periaqueductal gray (PAG), which is a major coordinator of emotional responses. The PAG acts via motor nuclei of the reticular formation and via the nuclei of cranial nerves, such as the nuclei of the vagus nerve. Another important subcortical site is the amygdala. The induction sites in the cerebral cortex, the cortical sites, include sectors of the anterior cingulate region and of the ventromedial prefrontal region* (Damasio, 1999, p. 60–61).
>
> (...)
>
> *The substrate for the representation of emotions is a collection of neural dispositions in a number of brain regions located largely in subcortical nuclei of the brain stem, hypothalamus, basal forebrain, and amygdala* (Damasio, 1999, p. 79).

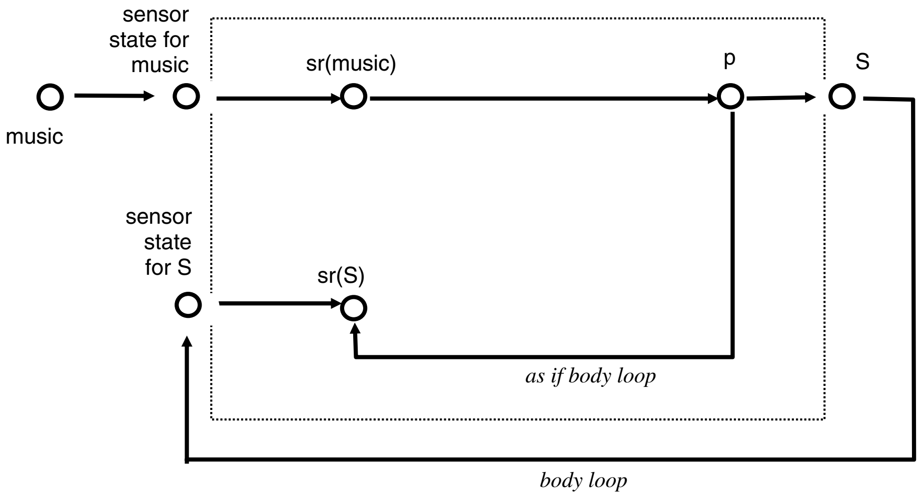

Fig. 2. Body loop and as-if body loop in the generation of feeling.

Moreover, for different types of emotions these sites are involved to varying degrees:

> *We have recently shown, using PET imaging, that the induction and experience of sadness, anger, fear, and happiness leand to activation in several of the sites mentioned above, but that the pattern for each emotion is distinctive. For instance, sadness consistently activates the ventromedial prefrontal cortex, hypothalamus, and brain stem, while anger or fear activate neither the prefrontal cortex nor hypothalamus. Brain stem activation is shared by all three emotions, but intense hypothalamic and ventromedial prefrontal activation appears specific to sadness* (Damasio, 1999, p. 60–61).

The model for emotions as described in this section abstracts from such specific processes. In particular, property LP2 could be refined and differentiated into a larger number of more specific causal relationships between neurological states involved in different types of emotions. However, as the main aim of the paper is to describe the global dynamics behind core consciousness as described in Damasio (1999), this more abstract and generic perspective was chosen.

The executable rules can also be used to specify functional roles of the internal state properties involved. The relational specification of the functional role of a mental state property concerns relationships both backward and forward in time. Given the model provided above, when looking backward, for example, the functional role of state property `p` is the temporal or causal relationship between `p` and the state that leads to `p`. Thus, looking backward the functional role of `p` is described by executable property LP2. Likewise, when looking forward the functional role of `p` is described by the causal relationship between `p` and the (future) state that is brought about by `p`, i.e., by executable property LP3.

4. Feeling

Next, Damasio's notion of *feeling* is considered. A central role is played by the proto-self that provides a map of the state of the organism's body:

> *'I propose that the sense of self has a preconscious biological precedent, the proto-self, and that the earliest and simplest manifestations of self emerge when the mechanism which generates core consciousness operates on that nonconscious precursor.*
> *The proto-self is a coherent collection of neural patterns which map, moment by moment, the state of the physical structure of the organism in its many dimensions. This ceaselessly maintained first-order collection of neural patterns occurs not in one brain place but in many, at a multiplicity of levels (...) These structures are intimately involved in the process of regulating the state of the organism. (...) We are not conscious of the proto-self.'* (Damasio, 1999, pp. 153–154).

Based on this proto-self he expresses the emergence of feeling as follows:

> *'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).

Next, he describes how the neural patterns which constitute the substrate of feeling arise in two classes of biological changes: changes related to body state and changes related to cognitive state. The former class of changes is briefly summarised as follows:

> *'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).

The changes related to cognitive state are briefly summarised by:

> *'They occur when the process of emotion leads to the secretion of certain chemical substances in nuclei of the basal forebrain, hypothalamus, and brain stem, and to the subsequent delivery of those substances to several other brain regions. When these nuclei release certain neuromodulators (such as monoamines) in the cerebral cortex, thalamus, and basal ganglia, they cause several significant alterations of brain function. The full range of alterations is not completely understood yet, but here are most important: (1) the induction of specific behaviors such as those aimed at generating bonding, nurturing, exploration, and playing; (2) a change in the ongoing processing of body states such that body signals may be filtered or allowed to pass, be selectively inhibited or enhanced, and their pleasant or unpleasant quality modified; and (3) a change in the mode of cognitive processing such that, for example, the rate of production of auditory or visual images can be changed (from slow to fast and vice versa) or the focus of images can be changed (from sharply focused to vaguely focused); changes in rate of production or focus are an integral part of emotions as disparate as those of sadness or elation.'* (Damasio, 1999, pp. 80).

Thus, a feeling emerges when the collection of neural patterns contributing to the emotion lead to mental images. In other words, the organism senses the consequences of the emotional state. In a generic manner, abstracting from the more specific and detailed biological states as described above, the two mechanisms by which a feeling can be achieved as distinguished by Damasio have been incorporated in the model:

(1) Via the *body loop*, the internal emotional state leads to a changed state of the body, which subsequently, after sensing, is represented in somatosensory structures of the central nervous system.
(2) Via the *as if body loop*, the state of the body is not changed. Instead, on the basis of the internal emotional state, a changed representation of the body is created directly in sensory body maps. Consequently, the organism experiences the same feeling as via the body loop: it is 'as if' the body had really been changed but it was not.

The model described in Section 3 has been extended to include a number of internal state properties for sensory representations of body state properties that are changed due to responses to the music; together these sensory representations constitute the feeling induced by the music. In Fig. 2 the conjunction of these sensory representations is depicted: sr(S) (a sensory representation of the changed body state; this may be materialised in a distributed manner as a kind of vector). This describes the 'body loop' for the responses on the music; here S and sensor_state(S) are effects and sensors in the body, respectively. In formal format, two additional local dynamic properties are needed (see also Fig. 2):

LP4 S ↠ sensor_state(S)
LP5 sensor_state(S) ↠ sr(S)

Notice that an internal state property sr(shivering) for shivering only, does not directly relate to the music. It is caused by the external stimulus shivering, which in this particular case is originally caused by the music. This body state property shivering could be present for a lot of other reasons as well, e.g., a cold shower. However, taking into account that not only shivering but a larger number of sensory state properties constitute the overall composite state property sr(S), the feeling will be more unique for the music. For the case of an 'as if body loop' dynamic properties LP3, LP4 and LP5 can be replaced by the following local dynamic property directly connecting p and sr(S).

LP6 p ↠ sr(S)

Also a combination of models can be made, in which some effects of hearing the music is caused by a body loop and some are caused by an 'as if body loop'.

5. Feeling a feeling

Finally, Damasio's notion of *knowing* or *being conscious of* or *feeling* a *feeling* is addressed. This notion is based on the organism detecting that its representation of its own (body) state (the *proto-self*) has been changed by the occurrence of a certain object (the music in our example). He expresses the way in which the proto-self contributes to a conscious feeling in the following hypothesis:

> *'Core consciousness occurs when the brain's representation devices generate an imaged, nonverbal account of how the organism's own state is affected by the organism's processing of an object, and when this process enhances the image of the causative object, thus placing it in a spatial and temporal context. (p. 169)... with the license of metaphor, one might say that the swift, second-order nonverbal account narrates a story: that of the organism caught in the act of representing its own changed state as it goes about representing something else. But the astonishing fact is that the knowable entity of the catcher has just been created in the narrative of the catching process. (...) You know it is you seeing because the story depicts a character – you – doing the seeing (pp. 170–172) ... beyond the many neural structures in which the causative object and the proto-self changes are separately represented, there is at least one other structure which re-represents both proto-self and object in their temporal relationship and thus represents what is actually happening to the organism: proto-self at the inaugural instant; object coming into sensory representation; changing of inaugural proto-self into proto-self modified by object.'* (*p. 177*).

In summary, the conscious feeling occurs when the organism detects the transitions between the following moments:

1. The proto-self exists at the inaugural instant.
2. An object comes into sensory representation.
3. The proto-self has become modified by the object.

For our case we restrict ourselves to placing the relevant events in a temporal context. In a detailed account, in the trace considered subsequently the following events take place: no sensory representations for music and S occur, the music is sensed, the sensory representation sr(music) is generated, the preparation representation p for S is generated, S occurs, S is sensed, the sensory representation sr(S) is generated. According to Damasio (1999, pp. 177–183), two transitions are relevant (see Damasio's Fig. 6.1, see also Fig. 3), and have to be taken into account in the model:

- from the sensory representation of the initial no S body state and not hearing the music to hearing music and a sensory representation of the music, and no S sensory representation;
- from a sensory representation of the music and no sensory representation of S to a sensory representation of S and a sensory representation of the music.

These two transitions are to be detected and represented by the organism. To model this process three internal state properties are introduced: s0 for encoding the initial situation, and s1 and s2 subsequently for encoding the situations after the two relevant changes. By making such state properties persistent they play the role of indicating that in the past a certain situation has occurred. With respect to the case study addressed, an example

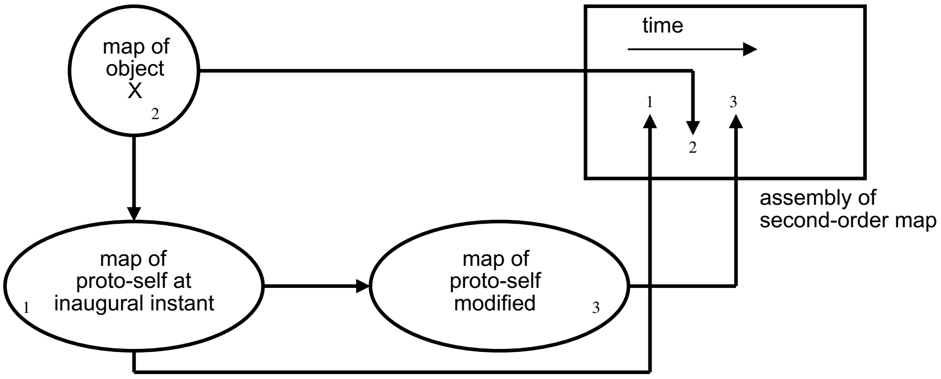
map of object X
2
time
1
2
3
assembly of second-order map
map of proto-self at inaugural instant
1
map of proto-self modified
3

Fig. 3. Damasio's picture for assembly of a second-order map.

behaviour, specified from an external perspective, is that the agent makes a statement about the music (e.g., "I am really touched by this music!"). Local dynamic properties that relate these additional internal state properties to the others can be expressed as follows (see also Fig. 4); here state properties s0 and s1 are persist

LP7 not sr(music) & not sr(S) ↠ s0
LP8 sr(music) & not sr(S) & s0 ↠ s1
LP9 sr(music) & sr(S) & s1 ↠ s2
LP10 s2 ↠ speak_about(music)

Given the model provided, when looking backward, the functional role of state property s2 for core consciousness is the causal relationship between s2 and the (past) states that cause s2. Thus, looking backward the functional role of s2 is described by executable property LP9. Likewise, when looking forward the functional role of s2 is described by the causal relationship between s2 and the (future) states that are caused by s2, i.e., by executable property LP10.

Damasio (1999, Ch. 8) addresses the neurological substrate for core consciousness. Evidence from patients with specific types of brain damage shows that core consciousness is disrupted in case of damage in one of: the cingulate gyrus, thalamic nuclei and superior colliculi. For the cingulate cortex, also referring to other relevant literature, he concludes:

> *'Reflection on the anatomical specifications of the cingulate cortex indicates that it is an excellent candidate for the sort of second-order structure I proposed earlier. Its different subregions and the massiveness of its somatosensory inputs can give rise to perhaps the most "integrated" view of the entire body state of an organism at any given time. But since the cingulate cortices are also privy to signals from the main sensory channels – the appearance of an object can be reported to the cingulate easily via both thalamic projections and direct projections from higher-order cortices in inferotemporal, polar temporal, and lateral parietal regions – the cingulate could help generate a neural pattern in which the relationship between the appearance of an object and the modifications undergone by the body could be mapped in the proper causal sequence.'* (Damasio, 1999, p. 264).

Referring to Stein and Meredith (1993), Damasio (1999) describes the role of the superior colliculi as follows:

> *'The superior colliculi are multilayered structures which receive a multiplicity of sensory inputs from an assortment of modalities, integrate signals in a complicated fashion across their several layers, and communicate the resulting outputs to a variety of brain-stem nuclei, the thalamus, and the cerebral cortex. (...) The integrative activity of the superior colliculi is aimed at orienting the eyes, the head and neck, and the ears (in creatures that move them) toward the source of a visual or auditory stimulus so that optimal object processing can take place. In the course of this activity, the superior colliculi map the temporal appearance*

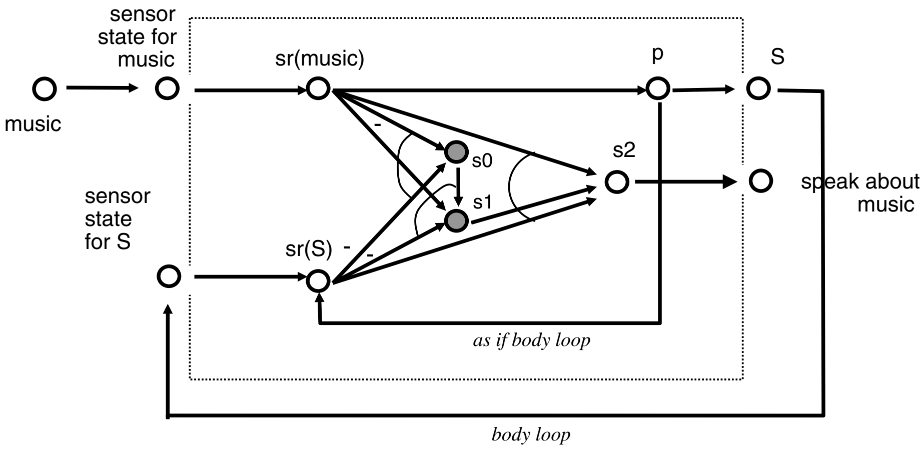

Fig. 4. Overview of the overall simulation model.

> *and spatial position of an object as well as varied aspects of body state. (...) In species with little cortical development this might be the source of the simple form of core consciousness that may accompany the execution of attentive behaviours. I hasten to add that, in the case of humans, there is no evidence that the superior colliculi can support core consciousness in the absence of the brain-stem and cingulate structures, even assuming intactness of the brain-stem proto-self structures.'* (Damasio, 1999, pp. 264–265).

In the way Damasio describes the role of the superior colliculi in core consciousness he does not go as far as Strehler (1994) who puts forward a more extreme view on them as 'the seat of consciousness'. The role of the thalamus is less well-documented, although 'bilateral damage to the thalamus disrupts consciousness for certain' (Damasio, 1999, p. 266).

6. Simulation

A special software environment has been created to enable the simulation of executable models (Bosse et al., 2007). Based on an input consisting of dynamic properties in LEADSTO format (and their timing parameters e, f, g, h, see Section 2), this software environment generates simulation traces. The algorithm used for the simulation is rather straightforward: at each time point, a bound part of the past of the trace (the maximum of all g values of all rules) determines the values of a bound range of the future trace (the maximum of $f + h$ over all LEADSTO rules). The software was written in SWI-Prolog/XPCE, and consists of approximately 20,000 lines of code. For more implementation details (see Bosse et al., 2007).

Using this software environment, the model described in the previous sections has been used to generate a number of simulation traces. Examples of such a simulation traces can be seen in Fig. 5(a) and (b). Here, time is on the horizontal axis, the state properties are on the vertical axis. A dark box on top of the line indicates that the property is true during that time period, and a lighter box below the line indicates that the property is false. This trace is based on all executable local properties (i.e., LP0 to LP9), except LP6. In all properties, the values (0, 0, 1, 1) have been chosen for the timing parameters e, f, g and h. Fig. 5(a) shows how the presence of the music first leads to an emotion (`p` or `S`), then to a feeling (`sr(S)`), and finally to the appearance of core consciousness (`s2`), involving a body loop.

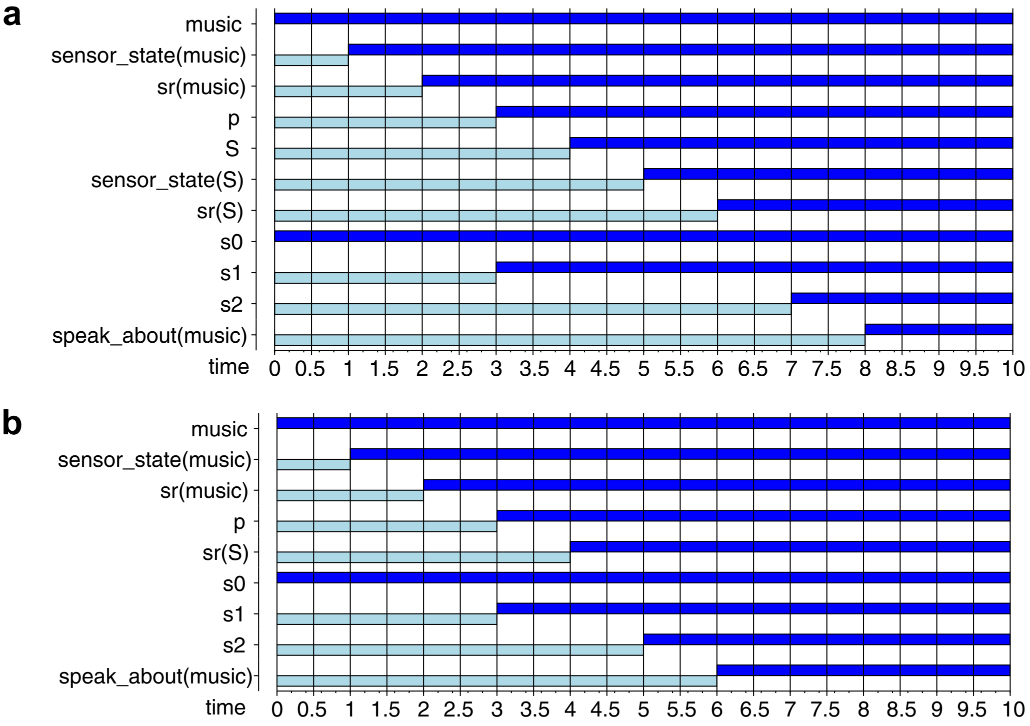

Fig. 5. Simulation trace involving (a) a body loop, and (b) an as-if body loop.

A similar trace is given in Fig. 5(b), for the case of the as-if body loop. This trace is based on all executable local properties (i.e., LP0 to LP9), except LP3, LP4 and LP5. Again, in all properties, the values $(0, 0, 1, 1)$ have been chosen for the timing parameters *e*, *f*, *g* and *h*. As can be seen in Fig. 5(b), in this case the feeling (sr(S)) immediately follows the preparatory state p, without an actual change in body state (S).

7. Representation relations

In this section, first in Section 7.1 it is discussed how in the literature in the area of consciousness the notion of representation is used in relation to dynamic and temporal aspects of the mental processes involved (e.g., Damasio, 1999; Dennett, 1991; Wegner, 2002). Next, in Section 7.2 some approaches to representational content of mental states from the literature on philosophy of mind are briefly discussed. In Sections 7.3–7.5 it is shown how these approaches can be applied to provide a philosophical and formal foundation of Damasio's use of the notion of representation (1999).

7.1. Uses of representation

The relationship between the notion of representation and dynamic or temporal aspects of mental processes attracts more and more attention in the literature (e.g., Dennett, 1991, 2001, 2005; Haselager, de Groot, & van Rappard, 2003; Pacherie, 2006; Pockett, Banks, & Gallagher, 2006; Wegner, 2002, 2003). A central theme here is that certain mental states or brain states can be interpreted as representing temporal information. Several authors put forward ideas on consciousness that incorporate this theme. It is not always clear, however, what is meant by the notion of representation used, and how it can be made more precise (e.g., Haselager et al., 2003). The literature in philosophy of mind provides a number of different approaches to define representational content; an interesting question is how these philosophical approaches relate to such literature on consciousness. In Damasio's description as addressed here various types of representation are mentioned, for example, sensory representations and second-order representations; more specifically, this can be seen in the following quotations (see also the previous ones).

> *'The substrate for the representation of emotions is a collection of neural dispositions in a number of brain regions (...) They exist, rather, as potential patterns of activity arising within neuron ensembles (...) On the one hand, the pattern of activation represents, within the brain, a particular emotion as 'neural object.'* (*p. 79*)
> *'... the body landscape is changed and is subsequently represented in somatosensory structures of the central nervous system, from the brain stem on up.'* (*p. 80*)
> *'Core consciousness occurs when the brain's representation devices generate an imaged, nonverbal account of how the organism's own state is affected by the organism's processing of an object'* (*p. 169*) *'... beyond the many neural structures in which the causative object and the proto-self changes are separately represented, there is at least one other structure which re-represents both proto-self and object in their temporal relationship and thus represent what is actually happening to the organism.'* (*p. 177*).

Also in other literature on consciousness representation of temporal relationships plays an important role. For example, Dennett (1991) discusses his multiple draft model, also called 'fame in the brain' or 'cerebral celebrity' model by Dennett (2001, 2005). A puzzling issue discussed by Dennett (1991, Ch. 5, 6), is the following. From a physical perspective within the brain an asynchronous distributed process occurs (in contrast, for example, to the nondistributed, synchronous process in a computer). A main question then is how the mind can create consciousness of events that are temporally ordered and can be synchronous. In this context he discusses how the brain can represent time (Ch. 6, Section 2).

As another example, also in the literature about conscious will such as by Pacherie (2006), Pockett et al. (2006) and Wegner (2002, 2003), representations of temporal relationships play an important role. Here, the perceived temporal relationship between a thought and an action is claimed to be a main source for the experience of ownership of an action. In a broader context, for example, in Kelley (1972, 1980) and Michotte (1954), representations for causal relationships are discussed.

In all of these cases, an interesting issue is how the notion of representation that is used can be analysed in the context of the philosophical approaches to representation found in the literature in philosophy of mind. In this paper the notion of representation as used by Damasio is analysed. It is shown how it can be made more precise using concepts from philosophy of mind and formalisations of them. In the remainder of this section, first three of these philosophical approaches are briefly introduced and subsequently it is discussed in how far the types of representation used by Damasio indeed can be related to these approaches.

7.2. Approaches to representational content

By Kim (1996, pp. 191–192) the *causal/correlational approach* to representational content is explained as follows. Suppose that, some causal chain is connecting an internal state property s and external state property 'horse nearby'. Due to this causal chain, under normal conditions internal state property s of an organism covaries regularly with the presence of a horse: this state property s occurs precisely when a horse is present nearby. Then the occurrence of s has the presence of the horse as its representational content. Especially for perceptual state properties this may work well.

By Kim (1996, pp. 200–202) the concept of *relational specification* of a state property is put forward as an other approach to representational content. It is based on a specification of how an internal state property can be related to properties of states distant in space and time.

> *'The third possibility is to consider beliefs to be wholly internal to the subjects who have them but consider their contents as giving relational specifications of the beliefs. On this view, beliefs may be neural states or other types of physical states of organisms and systems to which they are attributed. Contents, then, are viewed as ways of specifying these inner states; wide contents, then, are specifications in terms of, or under the constraints of, factors and conditions external to the subject, both physical and social, both current and historical.' (...)*
> *'The approach we have just sketched has much to recommend itself over the other two. It locates beliefs and other intentional states squarely within the subjects; they are internal states of the persons holding them, not something that somehow extrudes from them. This is a more elegant metaphysical picture than its alternatives. What is "wide" about these states is their specifications or descriptions, not the states themselves.'* (Kim, 1996, pp. 200–202).

In Kim's proposal a mental state property of a subject itself is distinguished from its relationships to other items. This contrasts to some other approaches where the mental state property is considered to be ontologically constituted as one entity comprising both the subject and the related items, or where the mental state property is considered to be the relation between the subject and the other items (cf. Kim, 1996, pp. 200–202). Kim explains how a mental state property itself can be considered an intrinsic internal state property, whereas its relational specification expresses how it relates to other items in the world. In particular, concentrating on the temporal dimension, a temporal relational specification can be viewed as the specification of temporal relationships of a (mental) state to other patterns in past and future. This approach is more liberal than the causal/correlational approach, since it is not restricted to one external state, but allows reference to a whole sequence of states in history.

More specifically, the idea is as follows. Suppose for an agent a mental state property p is given, which relates to a pattern of past (e.g., external world) traces (from a given time point t), on the one hand, and to a pattern of future (e.g., external world) traces on the other hand. Let $\varphi(\gamma, t)$ be a specification of this pattern of past traces γ and $\psi(\gamma, t)$ a specification of the pattern of future traces γ. Then the relational specification approach, which considers how the occurrence of mental state property p in the present relates to these past and future patterns, can be formalised as follows:

$$\varphi(\gamma, t) \iff \texttt{state}(\gamma, t) \models \texttt{p} \quad \texttt{(backward representation relation)}$$
$$\texttt{state}(\gamma, t) \models \texttt{p} \iff \psi(\gamma, t) \quad \texttt{(forward representation relation)}$$

Finally, in some recent literature cognitive functioning is studied from an interactivist perspective (e.g., Bickhard, 1993, 2000). The *temporal-interactivist approach* (Bickhard, 1993; Jonker & Treur, 2003) relates the occurrence of internal state properties to sets of past and future interaction traces. Bickhard (1993) emphasises the relation between the (mental) state of a system (or agent) and its past and future in the interaction with its environment as follows:

> *'When interaction is completed, the system will end in some one of its internal states - some of its possible final states. (...) The final state that the system ends up in, then, serves to implicitly categorise together that class of environments that would yield that final state if interacted with. (...) The overall system, with its possible final states, therefore, functions as a differentiator of environments, with the final states implicitly defining the differentiation categories. (...) Representational content is constituted as indications of potential further interactions.'*

Here it is indicated that mental states are related to interaction histories on the one hand, and to future interactions, on the other hand. Bickhard (1993, 2000) does not address the question how to formalise the interactivist approach, but in (Jonker & Treur, 2003) a formalisation is proposed which takes into account the temporal aspects of this interactivist perspective.

The general idea is as follows (cf. Jonker & Treur, 2003). Suppose for an agent a mental state property p is given, which relates to a pattern of past interaction traces (from a given time point t), on the one hand, and to a pattern of future interaction traces on the other hand. Let $\varphi(\gamma, t)$ be a specification of this pattern of past interaction traces γ and $\psi(\gamma, t)$ a specification of the pattern of future interaction traces γ. The temporal-interactivist approach considers the mental state property p holding in the present can mediate in this process as follows:

$$\varphi(\gamma, t) \Rightarrow \mathtt{state}(\gamma, t) \mid= \mathtt{p} \quad \& \quad \mathtt{state}(\gamma, t) \mid= \mathtt{p} \Rightarrow -\psi(\gamma, \mathtt{t})$$

Thus, like the relational specification approach, this approach allows reference to a whole sequence γ of states in history (or future). However, whilst in the relational specification approach these states can have any desired type (e.g., internal, external, or interaction states), in the temporal-interactivist approach they are restricted to interaction states (i.e., observations and actions).

The following sections explore whether these approaches can be used to specify the representational content of the relevant mental states that occur in our model (i.e., the states that represent emotion, feeling, and feeling a feeling). The focus is on the causal/correlational approach and the relational specification approach. The temporal-interactivist approach is not discussed in further detail. However, the formulae expressing the representational content according to the relational specification approach given below can be easily translated into the temporal-interactivist approach by replacing the external states that occur in the formulae by interaction states; e.g., replacing

```
    music                  by  sensor_state(music),
```
and
```
    speaks_about(music)    by  communicates (speaks_about(music))
```

7.3. Representational content of emotion

Consider the causal chain leading to a sequence of state property occurrences music, sensor_state(music), sr(music), p, S (see Fig. 1). Thus, looking backward in time, the external emotional state property S can be considered to (externally) represent the emotional content of the music. On the other hand, the internal emotional state property involved is p. Given the causal chain above the (backward) representational content for both p and S is the presence of this very special music, which could be considered acceptable. However, following the same causal chain, also the state property sr(music) has the same representational content. What is different between p and sr(music)? Why are the emotional responses to the same music different between different individuals? This would not be explainable if in all cases the same representational content is assigned. It might be assumed that state properties such as sr(music) may show changes between different individuals. However, the differences are probably much larger between the ways in which for two different individuals sr(music) is connected to a composite state property p. This subjective aspect is not taken into account in the causal/correlational approach. The content of such an emotional response apparently is more personal than a reference to an objective external factor, so to define this representational content both the external music and the internal personal make up has to be taken into account.

For the relational specification approach the representational content of p can be specified in a manner similar to the causal/correlational approach by ‘p occurs if the very special music just occurred’, and conversely. However, other, more suitable possibilities are available as well, such as, ‘p occurs if the very special music just occurred, and by this organism such music was perceived as sr(music) and for this organism sr(music) leads to p’, and conversely. This relational specification involves both the external music and the internal make up of the organism, and hence provides a subjective element in the representational content, in addition to the external reference. This provides an explanation of differences in emotional content of music between individuals.

7.4. Representational content of feeling

The representational content of sr(S) according to the causal/correlational approach can consider the causal chain music - sensor_state(music) - sr(music) - p - S - sensor_state(S) - sr(S). Using this chain, sr(S) can be related to both the presence of S, and further back to the presence of the very special music. This steps outside the context of having a reference to one state, which limits the causal/correlational approach. A more suitable approach is the relational specification approach, which allows such temporal relationships to different states in the past; there is the following temporal relation between the occurrence of sr(S), the presence of S, and the presence of music: ‘sr(S) occurs if S just occurred, preceded by the presence of the music’, and conversely:

$$\begin{aligned}
&\forall \mathrm{t1, t2}\ [\mathrm{t1} \leqslant \mathrm{t2}\ \&\ \mathrm{state}(\gamma, \mathrm{t1}, \mathrm{EW}) \models \neg \mathrm{S} \wedge \mathrm{music}\ \&\ \mathrm{state}(\gamma, \mathrm{t2}, \mathrm{EW}) \models \mathrm{S}]\\
&\quad \Rightarrow \exists \mathrm{t3} \geqslant \mathrm{t2}\ \mathrm{state}(\gamma, \mathrm{t3}, \mathrm{internal}) \models \mathrm{sr(S)}]\\
&\forall \mathrm{t3}\ [\mathrm{state}(\gamma, \mathrm{t3}, \mathrm{internal}) \models \mathrm{sr(S)}\\
&\quad \Rightarrow \exists \mathrm{t1, t2}\ \mathrm{t1} \leqslant \mathrm{t2} \leqslant \mathrm{t3}\ \&\ \mathrm{state}(\gamma, \mathrm{t1}, \mathrm{EW}) \models \neg \mathrm{S} \wedge \mathrm{music}\ \&\ \mathrm{state}(\gamma, \mathrm{t2}, \mathrm{EW}) \models \mathrm{S}]
\end{aligned}$$

7.5. Representational content of feeling a feeling

The backward representational content of s0 according to the causal/correlational approach can be taken as the absence of both S and music in the past, via the causal chain: no S and no music - sensor state no S and sensor state no music - no sr(music) and no sr(S) - s0. This can be expressed relationally by referring to one state in the past: ‘if no S and no music occur, then later s0 will occur,’ and conversely. Formally:

$$\begin{array}{lll}
\forall \mathrm{t1}\ [\mathrm{state}(\gamma, \mathrm{t1}, \mathrm{EW}) \models \neg \mathrm{S} \wedge \neg \mathrm{music} & \Rightarrow & \exists \mathrm{t2} \geqslant \mathrm{t1}\ \mathrm{state}(\gamma, \mathrm{t2}, \mathrm{internal}) \models \mathrm{s0}]\\
\forall \mathrm{t2}\ [\mathrm{state}(\gamma, \mathrm{t2}, \mathrm{internal}) \models \mathrm{s0} & \Rightarrow & \exists \mathrm{t1} \leqslant \mathrm{t2}\ \mathrm{state}(\gamma, \mathrm{t1}, \mathrm{EW}) \models \neg \mathrm{S} \wedge \neg \mathrm{music}]
\end{array}$$

For s1 and s2 the causal/correlational approach for backward representational content does not work well because these state properties essentially encode (short) histories of states. For example, the backward representational content of s1 according to causal/correlational approach can be tried as follows: presence of the music and no S in the past under the condition that at some point in time before that point in time no music occurred. However, this cannot be expressed adequately according to the causal/correlational approach since it is not one state in the past to which reference is made, but a history given by some temporal sequence. The problem is that no adequate solution is possible, since the internal state properties should in fact be related to sequences of different inputs over time in the past. This is something the causal/correlational approach cannot handle, as reference has to be made to another state at one time point, and it is not possible to refer to histories, i.e., sequences of states over time, in the past. A better option is provided by representational content of s1 as temporal relational specification: ‘if no S and no music occur, and later music occurs and still no S occurs, then still later s1 will occur,’ and conversely. Formally:

$$\begin{aligned}
&\forall \mathrm{t1, t2}\ [\mathrm{t1} \leqslant \mathrm{t2}\ \&\ \mathrm{state}(\gamma, \mathrm{t1}, \mathrm{EW}) \models \neg \mathrm{S} \wedge \neg\ \mathrm{music}\ \&\ \mathrm{state}(\gamma, \mathrm{t2}, \mathrm{EW}) \models \neg\ \mathrm{S} \wedge \mathrm{music}\\
&\quad \Rightarrow \exists \mathrm{t3} \geqslant \mathrm{t2}\ \mathrm{state}(\gamma, \mathrm{t3}, \mathrm{internal}) \models \mathrm{s1}]\\
&\forall \mathrm{t3}\ [\mathrm{state}(\gamma, \mathrm{t3}, \mathrm{internal}) \models \mathrm{s1}\\
&\quad \Rightarrow \exists \mathrm{t1, t2}\ \mathrm{t1} \leqslant \mathrm{t2} \leqslant \mathrm{t3}\ \&\ \mathrm{state}(\gamma, \mathrm{t1}, \mathrm{EW}) \models \neg \mathrm{S} \wedge \neg \mathrm{music}\ \&\ \mathrm{state}(\gamma, \mathrm{t2}, \mathrm{EW}) \models \neg \mathrm{S} \wedge \mathrm{music}]
\end{aligned}$$

Similarly, the backward representational content of s2 as relational specification can be specified as follows: ‘if no S and no music occur, and later music occurs and still no S occurs, and later music occurs and S occurs, then still later s2 will occur,’ and conversely. Formally:

```
∀t1, t2, t3 [t1⩽ t2 ⩽ t3 &
   state(γ, t1, EW) |= ¬ S ∧ ¬ music &
   state(γ, t2, EW) |= ¬ S ∧ music &
   state(γ, t3, EW) |= S ∧ music ⇒
      ∃t4 ⩾ t3 state(γ, t4, internal) |=s2]


∀t4 [state(γ, t4, internal) |= s2 ⇒
   ∃t1, t2, t3 t1⩽ t2 ⩽ t3 ⩽ t4 &
   state(γ, t1, EW) |= ¬ S ∧ ¬music &
   state(γ, t2, EW) |= ¬ S ∧ music &
   state(γ, t3, EW) |= S ∧ music]
```

This comes close to the transitions indicated by Damasio (1999, p.177), as also mentioned in Section 5: *the proto-self exists at the inaugural instant—an object comes into sensory representation—the proto-self has become modified by the object*.

The above relational specification is a first-order representation in the sense that it refers to external states of world and body, whereas Damasio's second-order representation refers to internal states (other, first-order, representations) of the proto-self. Moreover, the relational specification given above only works for body loops, not for 'as if body loops'. A relational specification that comes more close to Damasio's formulation, and also works for 'as if body loops' is the following (**RSP**):

```
∀t1, t2, t3 [t1⩽ t2 ⩽ t3 &
   state(γ, t1, internal) |= ¬sr(S) ∧ ¬sr(music) &
   state(γ, t2, internal) |= ¬sr(S) ∧ sr(music) &
   state(γ, t3, internal) |= sr(S) ∧ sr(music) ⇒
      ∃t4 ⩾ t3 state(γ, t4, internal) |= s2]


∀t4 [state(γ, t4, internal) |= s2 ⇒
   ∃t1, t2, t3 t1⩽ t2 ⩽ t3 ⩽ t4 &
   state(γ, t1, internal) |= ¬sr(S) ∧ ¬sr(music) &
   state(γ, t2, internal) |= ¬sr(S) ∧ sr(music) &
   state(γ, t3, internal) |= sr(S) ∧sr(music)]
```

This is a relational specification in terms of other representations (i.e., sr(music), sr(S)), and therefore a second-order representation. It has no direct reference to external states anymore. However, indirectly, via the first-order representations sr(music) and sr(S) it has references to external states.

When looking forward, the representational content of mental state property s2 can be described by relating it to future world states. For the given model the forward representational content of state property s2 can be informally described as follows: 'if s2 occurs, then later the agent will speak about the music', and conversely. This expression is formalised as follows:

```
∀t1 [state(γ, t1, internal) |= s2 ⇒ ∃t2 ⩾ t1 state(γ, t2, EW) |=speaks_about(music)]
∀t2 [state(γ, t2, EW) |= speaks_about(music) ⇒ ∃t1 ⩽ t2 state(γ,t1,internal)|=s2]
```

8. Verification

In Sections 3–6, local, executable dynamic properties were addressed, and simulation based on these properties was discussed. In Section 7, dynamic properties to describe representational content of internal states were introduced. These dynamic properties are of a *global* nature. Another example of a more global property is the following:

OP1 music ↠ s2

Informally, this property states that the presence of music eventually leads to the birth of core consciousness (s2). This can be considered as a global property because it describes dynamics of the overall process, whereas the local properties (LP's) presented in Sections 3–6 described basic steps of the process.

In principle, various behaviours can be described in which core consciousness fulfils a role. Examples in general are high-level cognitive functions such as reasoning, planning and language processing. With respect to the above case study, an example behaviour, specified from an external perspective, is that the agent makes a statement about the music (e.g., "I am really touched by this music!"). Using the executable format, this behaviour can be formalised by the following global behavioural property:

GP1 music $\twoheadrightarrow$ speak_about(music)

For all types of global properties (i.e., dynamic properties OP1 and GP1 and the properties specifying representational content), an important issue is *verification*. In other words, are these global properties satisfied by the simulation model described in Sections 3–6? Therefore, the global properties have been formalised, and verification has been applied in two ways: by *automated checks* on simulation traces and by establishing and automatically verifying *logical relationships*.

8.1. Automated checks on simulated traces

In addition to the simulation software described in Section 6, a software environment has been developed that enables one to check dynamic properties specified in TTL against simulation traces. This software environment takes a dynamic property and one or more (empirical or simulated) traces as input, and checks whether the dynamic property holds for the traces. Using this environment, the global properties mentioned above have been automatically checked against traces like depicted in Figs. 5(a),(b). The duration of these checks varied between 0.5 and 1.5 s, depending on the complexity of the formula. All these checks turned out to be successful, which validates (for the given traces at least) our choice for the representational content of the internal state properties. However, note that these checks are only partial validation, they are no exhaustive proof of validities (statements proven true on *all* possible traces) as, e.g., model checking (Clarke, Grumberg, & Peled, 1999; McMillan, 1993) is; see, however, next subsection.

8.2. Verifying logical relationships

A second way of verification is to establish logical relationships between global properties and local properties. This has been performed in a number of cases. For example, to relate OP1 to local properties, intermediate properties were identified in the form of the following milestone properties that split up the process in three phases:

MP1(MtoE)	music $\twoheadrightarrow$ sr(music) & sr(music) $\twoheadrightarrow$S
MP2(EtoF)	S $\twoheadrightarrow$ sr(S)
MP3(FtoFF)	**RSP** (see Section 7)

For the milestone properties the following relationships hold (for simplicity neglecting 'as if body loops'):

MP1(MtoE) & MP2(EtoF) & MP3(FtoFF)	⇒ OP1
LP0 & LP1 & LP2 & LP3	⇒ MP1(MtoE)
LP4 & LP5	⇒ MP2(EtoF)
LP7 & LP8 & LP9	⇒ MP3(FtoFF)

Fig. 6 provides the same relationships in the form of a logical AND-tree.

Such logical relationships between properties can also be very useful in the analysis of traces. For example, if a given trace that is unsuccessful does not satisfy milestone property MP2, then by a refutation process it can be concluded that the cause can be found in either LP4 or LP5. In other words, either the sensor mechanism fails (LP4), or the sensory representation mechanism fails (LP5). Using the model checking environment SMV (cf. McMillan, 1993), it has been automatically verified that indeed the local properties LP1 through LP9

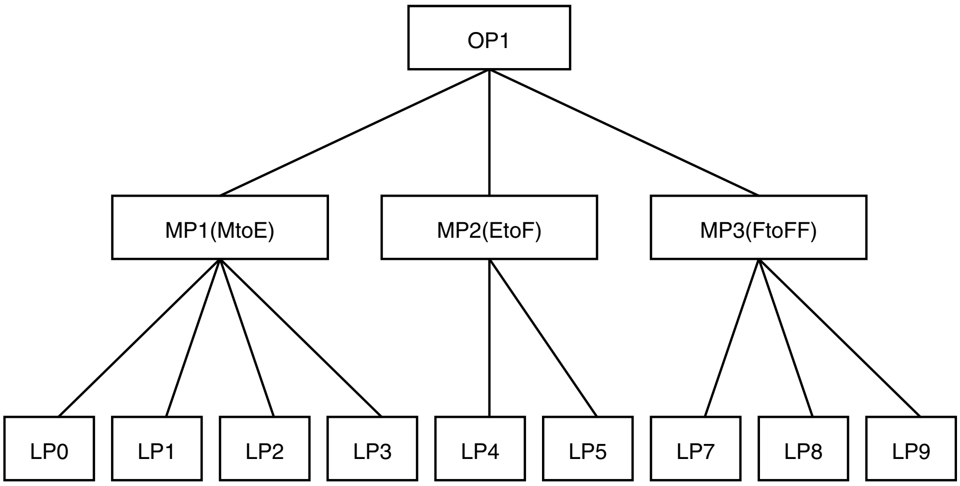

Fig. 6. Logical relationships between dynamic properties.

together entail global property OP1. Likewise, it has been verified that these local properties together entail the representational content specifications. This approach based on different aggregation levels of a process and interlevel relations has similarities with the notion of componential explanation as described by Clark (1997), Cummins (1975, 1983) and Davies (2001) (see also Bosse, Jonker, & Treur, 2006a).

9. Discussion

The aim of the work reported here was to analyse in how far the theory of core consciousness described by Damasio (1999) can be formalised in a coherent manner. The outcome of the work shows that indeed this is the case. The indications about relations between different states lead to a suitable computational model of the basic mechanisms of the process that shows behaviour as expected. Moreover, the statements made in Damasio (1999) about representations also can be formalised, and fit the other part of the formalisation in the sense that these are indeed logically entailed by the computational model of the basic mechanisms. Following Damasio (1999), the chosen modelling approach describes temporal dependencies in processes at a neurological, rather than symbolic level. To avoid complexity the model was specified at an abstract level. More specifically, from the available approaches to representational content from philosophy of mind, the causal/correlational approach is not applicable, but Kim's relational specification approach (1996), that allows more complex temporal dependencies, is applicable. Using this approach, claims on representational content made by Damasio (1999) have been formalised and supported using two types of techniques for verification against the computational model describing the basic mechanisms. The concepts used in the model closely follow the informally described concepts by Damasio (1999). Further discussion of the relationships between these concepts and neuroanatomical structures has been described by Damasio (1999, Ch. 8, pp. 234–276), amongst others. As these suggested relationships are still hypothetical to a certain extent, research is still needed to fully validate them; for partial confirmations (see, for example, Damasio et al., 2000; Parvizi & Damasio, 2001; Parvizi et al., 2006).

Furthermore, while developing the computational model, by the authors a prediction has been made on the basis of the formalisation, namely the possibility of 'false core consciousness': core consciousness that is attributed to the 'wrong' stimulus. To explain this phenomenon, suppose that two stimuli occur, say x1 and x2, where x2 is subliminal and unnoticed. Then, it could be the case that x2 provokes emotional responses, whilst the conscious feeling that arises is attributed to x1 instead of x2. In terms of the model presented here, this can be simulated by first introducing a subliminal stimulus that yields emotion S (e.g., a cold breeze) followed by the stimulus music. In that case, the conscious feeling would incorrectly be attributed to the music. Putting this prediction forward in email communication to Damasio, he confirmed the existence of this predicted false core consciousness. See also the extensive discussion of empirical findings on this issue in Prinz (2004, Ch. 3).

Although this is not a proof for the validity of the model, it indicates that this type of modelling can be used to derive interesting predictions.

From a philosophical perspective the paper contributes a case study for representational content which is more down-to-Earth (and more complex) than the science fiction style thought experiments, such as the planet Twin Earth, that are common in the literature on philosophy of mind (e.g., Kim, 1996). In addition, the type of representation is more sophisticated than the usual ones essentially addressing sensory representations induced by observing (a snapshot of) a horse or a tomato. Interesting further work in this area is to analyse other approaches given in this literature on representational content by applying them to this example.

As discussed in Section 7.1, also in other literature representation of temporal relationships plays an important role. For example, in the context of the multiple draft model, Dennett (1991) discusses how representations of time may allow the brain, as an asynchronous distributed process, to create conscious experience of temporally ordered events (Ch. 6, Section 2). It may be an interesting study to analyse how such representations can be made more precise in a similar way as addressed for Damasio's notions of representation. The same can be said for the literature about conscious will such as papers by Pacherie (2006), Pockett et al. (2006) and Wegner (2002, 2003). Here, representations of temporal relationships between a thought and an action are considered a basis for the experience of ownership of an action. Also this may be analysed using similar methods. Two papers reporting other analyses of more complex cases of representational content are by Bosse, Jonker, and Treur (2005a) for representational content for a case of intensive reciprocal agent-environment interaction, and Bosse, Jonker, Schut, and Treur (2006) for collective representational content for a case of a society of agents.

As indicated in the introduction, the effort to obtain a formalisation of an informally described theory can only be justified if certain gains can be recognised. The items (a) to (g) mentioned in the introduction are evaluated as follows.

(a) Assessing the theory with respect to possible unintended conceptual incompleteness or ambiguity, and if present, identification (and possibly removal) of such ambiguities.

It has turned out that the basic conceptual framework has been described by Damasio (1999) in such a manner that on the crucial issues no incompleteness or ambiguity was found to stand in the way of a faithful formalisation. In experiences with other informally formulated texts sometimes a serious step of conceptual clarification and precision had to made before formalisation was possible. In this case this was not needed, as Damasio (1999) already contributed that step. As an example, the notion of second-order representation which is the essence of the theory on core consciousness is described in so much detail and clarity that the formalisation of the representation relation given in Section 7 above, almost literally follows the informal formulation by Damasio (1999). In this way the formalisation effort contributes a positive assessment for the theory as described.

(b) Assessing the theory with respect to internal coherency or consistency, and if present, detection (and possibly removal) of possible incoherencies or inconsistencies.

In addition to the above, during the modelling and formalisation also no incoherencies or inconsistencies have been found. For example, the notions of first-order and second-order representation used in the theory can be justified by an approach in the literature on philosophy of mind; see also item (e). In this way another positive assessment for the theory is contributed.

(c) Increasing detailedness of the theory

By the formalisation, some increase in detailedness has been obtained, for example, in the distinction between the state properties `s0`, `s1 and s2` and the precise temporal relations between them. However, in the light of the positive assessment of (a), the increase in detailedness was limited.

(d) Deriving more detailed implications of the theory

One of the implications of the theory found was the notion of false core consciousness discussed above. This implication was found once the model was designed and it became clear what the possibilities and impossibilities of this model are.

(e) Relating the theory to other theories and perspectives in the literature

An interesting question that was addressed by the formalisation effort was in how far the use of the notion of representation by Damasio (1999) can be justified by approaches to representation in the literature on philosophy of mind. By relating this notion to Kim (1996)'s notion of relational specification of representational content and formalising this notion, this question has been answered in a positive and even formally founded manner. This is a substantial contribution of the formalisation effort to the theory. Even when the precise formalisation is left out of consideration, the underlying conceptual analysis is an interesting contribution providing a cross-connection between this theory as described in the literature in neuroscience or cognitive science, and the literature in philosophy of mind.

(f) The possibility to conduct pseudo-experiments (simulations) for the theory by computer support

The theory was formalised in a computational manner. The simulations performed indeed show how pseudo-experiments can be conducted. The model can be extended by additional aspects concerning specific experimental setups (e.g., involving multiple stimuli, or repetition of a stimulus, or introducing specific physical factors that disturb the body state, such as a cold shower or a certain medicine) to obtain simulated outcomes within a certain experimental context, after which simulation results can be compared with results of the corresponding experiments.

(g) The possibility to use computer support for verification and validation of the theory

In the first place the formalisation provides the possibility to perform automated verification of properties on simulated or empirical process traces. The formalised state ontologies can be used as a format to represent the states in either type of traces in the computer. In the case of simulated traces this can be done by an automated translation. In the case of empirical traces, the formalisation of the trace into a format readable by the computer is done by hand. But once such a formal representation is available, any dynamic property can be checked for it. A second way in which automated verification is offered is in relation to interlevel relations.

The analysis approach that is applied in this paper to model Damasio's theory of consciousness, has previously been applied to complex and dynamic cognitive processes other than consciousness, such as the interaction between agent and environment (Bosse et al., 2005a). In a number of these cases, in addition to simulated traces, also empirical (human) traces have been formally analysed. Using this approach, it is possible to verify global dynamic properties (e.g., specifying the representational content of internal states) in real-world situations.

For some other recent work in the area of emotion and consciousness (see Prinz, 2004, Ch. 3), which gives an account for emotions as embodied representations of 'core relational themes' such as danger and obstruction. An interesting extension of the work described here would be to make a similar formal analysis of this perspective, and to compare this formalisation with the formalisation of Damasio's theory. The same could be done with work reported by Edelman and Tononi (2000) and Metzinger (2003).

Acknowledgments

The authors are grateful to Antonio Damasio for constructive and stimulating feedback on this work in different stages. The work has also had benefit from discussions at the AISB'05 Symposium on Machine Consciousness, and at the ICCM'06 conference on cognitive modelling, where the work has been presented in a preliminary form (cf. Bosse, Jonker, & Treur, 2005b, 2006b). The current paper is a more worked out extension of this earlier work.

References

Bickhard, M. H. (1993). Representational content in humans and machines. *Journal of Experimental and Theoretical Artificial Intelligence, 5*, 285–333.

Bickhard, M. H. (2000). Information and representation in autonomous agents. *Journal of Cognitive Systems Research, 1*, 65–75.

Bosse, T., Jonker, C. M., van der Meij, L., & Treur, J. (2007). LEADSTO: A language and environment for analysis of dynamics by simulation. *International Journal of Artificial Intelligence Tools, 16*, 435–464.

Bosse, T., Jonker, C. M., Schut, M. C., & Treur, J. (2006). Collective representational content for shared extended mind. *Cognitive Systems Research Journal, 7*, 151–174.

Bosse, T., Jonker, C.M., Treur, J. (2005a). Representational content and the reciprocal interplay of agent and environment. In J. Leite, A. Omincini, P. Torroni, & P. Yolum (Eds.), *Proceedings of the second international workshop on declarative agent languages and technologies, DALT'04. Lecture Notes in AI* (Vol. 3476, pp. 270–288). Springer-Verlag.

Bosse, T., Jonker, C. M., & Treur, J. (2005b). Simulation and representation of body, emotion, and core consciousness. In *Proceedings of the AISB 2005 symposium on next generation approaches to machine consciousness: Imagination, development, intersubjectivity, and embodiment* (pp. 95–103). London: AISB Publishers.

Bosse, T., Jonker, C.M., & Treur, J., (2006a). Componential explanation in philosophy, cognitive science and computer science. In: R. Sun, N. Miyake (Eds.), *Proc. of the 28th annual conference of the cognitive science society, CogSci'06* (pp. 95–100).

Bosse, T., Jonker, C.M., & Treur, J., (2006b). Formal Analysis of Damasio's Theory on Core Consciousness. In: D. Fum, F. Del Missier, & A. Stocco (Eds.), *Proceedings of the seventh international conference on cognitive modelling, ICCM'06* (pp. 68–73). Edizioni Goliardiche.

Clark, A. (1997). *Being there: Putting brain, body and world together again*. Cambridge Massachusetts, London England: MIT Press.

Clarke, E. M., Grumberg, O., & Peled, D. A. (1999). *Model checking*. Cambridge Massachusetts, London England: MIT Press.

Cummins, R. (1975). Functional analysis. *The Journal of Philosophy, 72*, 741–760.

Cummins, R. (1983). *The nature of psychological explanation*. Cambridge, Mass: MIT Press.

Damasio, A. (1999). The feeling of what happens: Body, emotion and the making of consciousness. *Harcourt Brace*.

Damasio, A., Grabowski, T. J., Bechara, A., Damasio, H., Ponto, L. L. B., Parvizi, J., et al. (2000). Subcortical and cortical brain activity during the feeling of self-generated emotions. *Nature Neuroscience, 3*, 1049–1056.

Davies, P. S. (2001). *Norms of nature: Naturalism and the nature of functions*. Cambridge, Mass: MIT Press.

Dennett, D.C. (1991). *Consciousness explained*. Little, Brown & Company. Penguin Books, 1993.

Dennett, D. C. (2001). Are we explaining consciousness yet? *Cognition, 79*, 221–237.

Dennett, D. C. (2005). *Sweet dreams: Philosophical obstacles to a science of consciousness*. Cambridge Massachusetts, London England: MIT Press.

Edelman, G. M., & Tononi, G. (2000). *A universe of consciousness*. Basic Books.

Galton, A. (2003). *Temporal Logic. Stanford Encyclopedia of Philosophy*. Available from URL: http://plato.stanford.edu/entries/logic-temporal/#2.

Galton, A. (2006). Operators vs arguments: The ins and outs of reification. *Synthese, 150*, 415–441.

Haselager, W. F. G., de Groot, A. D., & van Rappard, J. F. H. (2003). Representationalism versus anti-representationalism: A debate for the sake of appearance. *Philosophical Psychology, 16*(1), 5–23.

Jonker, C. M., & Treur, J. (2002). Compositional verification of multi-agent systems: A formal analysis of pro-activeness and reactiveness. *International Journal of Cooperative Information Systems, 11*, 51–92.

Jonker, C. M., & Treur, J. (2003). Temporal-interactivist perspective on the dynamics of mental states. *Cognitive Systems Research Journal, 4*, 137–155.

Jonker, C. M., Treur, J., & Wijngaards, W. C. A. (2003). A temporal modelling environment for internally grounded beliefs, desires and intentions. *Cognitive Systems Research Journal, 4*, 191–210.

Kelley, H. H. (1972). Attribution in social interaction. In E. E. Jones, D. E. Kanouse, H. H. Kelley, R. E. Nisbett, S. Valins, & B. Weiner (Eds.), *Attribution: Perceiving the causes of behavior* (pp. 1–26). Morristown NJ: General Learnbing Press.

Kelley, H. H. (1980), Magic tricks: The management of causal attributions. In: D. Goerlitz (Ed.), *Perspectives on attribution research and theory: The bielefeld symposium* (pp. 19–35). Cambridge MA: Ballinger.

Kim, J. (1996). *Philosophy of mind*. England Oxford: Westview Press.

McMillan, K.L. (1993). *Symbolic model checking: An approach to the state explosion problem*. PhD thesis, School of computer science, Carnegie Mellon University, Pittsburgh, 1992. Kluwer Academic Publishers.

Metzinger, Th. (2003). *Being no one. The self-model theory of subjectivity*. Cambridge MA: MIT Press.

Michotte, A. (1954). *The perception of causality* (T.R. Miles & E. Miles, Trans.). New York: Basic Books (1963).

Pacherie, E. (2006). Toward a dynamic theory of intentions. In S. Pockett, W. P. Banks, & S. Gallagher (Eds.), *Does consciousness cause behaviour?* (pp 144–167). Cambridge Massachusetts, London England: MIT Press.

Parvizi, J., & Damasio, A. (2001). Consciousness and the brain stem. *Cognition, 79*, 135–159.

Parvizi, J., Hoesen, G. W., van Buckwalter, J., & Damasio, A. (2006). Neural connections of the posteromedial cortex in the macaque: Implications for the understanding of the neural basis of consciousness. *Proceedings of National Academy of Sciences, 103*(5), 1563–1568.

Pockett, S., Banks, W. P., & Gallagher, S. (Eds.). (2006). *Does consciousness cause behaviour?* Cambridge Massachusetts, London England: MIT Press.

Prinz, J. J. (2004). *Gut reactions: A perceptual theory of emotion*. Oxford England: Oxford University Press.

Reiter, R. (2001). *Knowledge in action: Logical foundations for specifying and implementing dynamical systems*. Cambridge Massachusetts, London England: MIT Press.

Stein, B. F., & Meredith, M. A. (1993). *The merging of the senses*. Cambridge Massachusetts, London England: MIT Press.

Strehler, B. (1994). Where is the self? A neuroanatomical theory of consciousness. *Synapse, 7*, 44–91.

Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge Massachusetts, London England: MIT Press.

Wegner, D. M. (2003). The mind's best trick: How we experience conscious will. *Trends in Cognitive Science, 7*, 65–69.