How the Brain Creates Emergent Information by the Development of Mental Models

Treur, Jan

published in
Mental Models and their Dynamics, Adaptation and Control
2022

DOI (link to publisher)
10.1007/978-3-030-85821-6_16

document version
Publisher's PDF, also known as Version of record

Chapter 16
How the Brain Creates Emergent Information by the Development of Mental Models: An Analysis from the Perspective of Temporal Factorisation and Criterial Causation

Jan Treur

Abstract Mental models usually are considered to represent a form of information or knowledge about situations or processes in the world or about mental states and processes in humans. In this chapter it is explored how this informational view on mental models relates to two more general principles that have been introduced in recent years: temporal factorisation and criterial causation. Propagated activation of neurons in the brain through their network co-occurs with adaptation of characteristics of this network such as connection strengths or excitability thresholds. According to the principle of criterial causation, these adaptive network characteristics can be interpreted as informational criteria for future activation of a considered neuron. This then is viewed as a form of emergent information formation and use of this information: the activation of neurons as determined by such information is termed criterial causation, where the information defines the criteria. The criterial causation principle strongly relates to the principle of temporal factorisation for the dynamics of the world in general, describing how the world represents information about its past in its present state, which then in turn determines the world's future. In the chapter, these processes are analysed in more detail and modeled by (adaptive) network models showing how mental models that are learnt can be interpreted as emerging information according to the principles of criterial causation and temporal factorisation.

Keywords Mental model · Criterial causation · Temporal factorisation · Adaptive network model · Emerging informational content

J. Treur (✉)
Social AI Group, Department of Computer Science , Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
e-mail: j.treur@vu.nl

J. Treur and L. Van Ments (eds.), *Mental Models and Their Dynamics, Adaptation, and Control*, Studies in Systems, Decision and Control 394,
https://doi.org/10.1007/978-3-030-85821-6_16

16.1 Introduction

Mental models can be viewed as a way to represent a form of information or knowledge about situations and processes in the world or about mental states and processes in humans. It turns out that this informational view on mental models relates to two more general principles for dynamics and information that have been introduced in recent years: temporal factorisation and criterial causation (Treur 2007a, b, 2021; Tse 2013, 2018). These relations for mental models are explored in some detail in the current chapter.

Neural processing is often considered as based on propagation of activation of neurons within the network they form; e.g., (Hebb 1949; Tse 2013). However, propagation within this network depends on characteristics of the network, under which connection strengths and the excitability thresholds of neurons. This makes a more complete picture of neural processes, as both dynamics *within* the network of neurons and dynamics *of* this network. Tse (2013, 2018) considers these settings that form a brain configuration as serving as criteria for the incoming signals to get a neuron activated; this is called *criterial causation*. The criteria are viewed as a form of *information* emerging in the brain, called by Tse (2013), p. 259 'physically realised informational criteria'; when in the future these criteria are fulfilled, the neuron will activate.

So, it is not only activation of neurons that changes over time, also network characteristics representing the criteria change over time. These changes are influenced by incoming patterns in the past. This change can be slow or fast in comparison to propagation of activation of neurons: slow as for learning from multiple experiences, or fast as in formation of memories, which is termed '*rapid resetting* of these criteria' (Tse 2013, 2018). This form of network adaptation creates *emerging information*, and activation of neurons is determined by the created information (Tse 2013, 2018).

The criterial causation principle can be considered a special case of the temporal factorisation principle which addresses the world dynamics more in general (Treur 2007a, b). Mediating state configurations encode information (on the past of the world) in the present world state, and they determine the future world patterns. The temporal factorisation principle generalises criterial causation to emerging formation of information and usage of this information as a mechanism of the world dynamics in general. Viewed in this way, the criterial causation principle for the brain makes use of that more general principle. The relation between the two is then that the temporal factorisation principle describes how the world dynamics creates and exploits emerging formation of information in its state, and criterial causation addresses particularly how the brain generates and exploits emerging information in its state. In each of these two cases this emergent information determines future patterns.

These two principles on dynamics and related emerging information are applied here to describe how a mental model that is learnt represents information defined by criteria for criterial causation or by a mediating state configuration for temporal factorisation. In this way learning a mental model can be viewed as an emergent process creating information and simulating the mental model as using this created information. This emerging information is what the mental model represents. This view provides an informational view on the learning and use of mental models.

In this chapter, Sect. 16.2 makes a more detailed analysis of temporal factorisation and criterial causation and how they relate. Section 16.3 briefly introduces the network-oriented modeling approach used in the chapter. Next, network models are presented to illustrate both principles. In Sect. 16.4, examples of temporal factorisation are addressed by (nonadaptive) network modeling, including results of an example simulation. Section 16.5 addresses how for networks criteria can be specified for criterial causation, and Sect. 16.6 presents an adaptive network model for criterial causation; in Sect. 16.7 results of an example simulation are discussed. In Sect. 16.8 it is discussed how the informational content of a mediating state for temporal factorisation and of emerging criteria for criterial causation can be defined, based on the concept of temporal relational specification from Philosophy of Mind. After that, in Sect. 16.9, for more technically interested readers some formal details are described which for the sake of readability were omitted in all previous sections. Section 16.10 is a discussion.

16.2 Temporal Factorisation Versus Criterial Causation

In this section the notions of Temporal Factorisation (Sect. 16.2.1) and Criterial Causation (Sect. 16.2.2) are introduced, and their relationship is discussed (Sect. 16.2.3).

16.2.1 Temporal Factorisation

The principle of *temporal factorisation* (Treur 2007a, b) expresses that if a temporal relationship $a \Rightarrow b$ from past pattern a to future pattern b holds, then this can be factorised by a state property p (indicating a configuration in the world) into temporal relationships $a \Rightarrow p$ from the past pattern a to present state p and $p \Rightarrow b$ from present state p to future pattern b. State property p is termed a *mediating state property* or *mediating state configuration* for $a \Rightarrow b$. For an illustration, see Fig. 16.1. Note that the principle expresses that for any such a relationship $a \Rightarrow b$ a world state property p *exists* such that the temporal relationships $a \Rightarrow p$ and $p \Rightarrow b$ hold:

$$[a \Rightarrow b] \Rightarrow \quad \exists p \quad [a \Rightarrow p \quad \& \quad p \Rightarrow b] \tag{16.1}$$

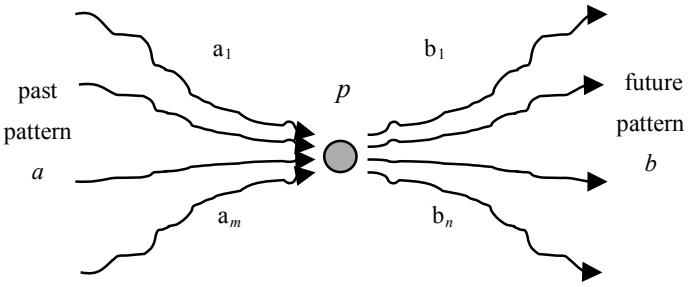


Fig. 16.1 Temporal factorisation: mediating state property p for the temporal past-to-future relationship $a \Rightarrow b$; picture adopted from Treur (2007a, p. 60, Fig. 1)

As mentioned, this mediating state property p describes a configuration of the world state in the present. It can (and often will) be not one simple state property, but a combination of a number of state properties. The notion of pattern is kept a bit informal here (however, see Sect. 16.9 for formalisations), but it can be understood as a property of possible temporal traces, where a temporal trace is a sequence of states for all time points. For example, in Fig. 16.1 pattern a is illustrated for specific traces a_1 to a_m fulfilling that pattern, and pattern b for traces $b_1, \ldots, b_n$. Such a property of temporal traces can be expressed as some temporal statement; see Sects. 16.8 and 16.9 for more explicit descriptions and examples.

The word factorisation is explained algebraically as follows. Suppose the set of possible past patterns are indicated by the set PP, the set of possible future patterns by the set FP, and the set of possible present states by PS. Then the temporal relationship $a \Rightarrow b$ can be described by an operator or function f, assigning future patterns b to past patterns a

$$f : \mathrm{PP} \rightarrow \mathrm{FP} \tag{16.2}$$

Temporal factorisation expresses that there are functions g and h with

$$\begin{aligned} g &: \mathrm{PS} \rightarrow \mathrm{FP} \\ h &: \mathrm{PP} \rightarrow \mathrm{PS} \end{aligned} \tag{16.3}$$

such that function f is factorised by operator g and operator h as follows:

$$f = g_0 h : \quad \mathrm{PP} \rightarrow \mathrm{PS} \rightarrow \mathrm{FP} \tag{16.4}$$

where $g_0 h$ denotes the functional composition of functions g and h: first h is applied then g.

According to the temporal factorisation principle, based on the past the present world state configuration p encodes all relevant information; then the world can ignore the past if it generates a future temporal pattern. So, it essentially claims that world state configurations from an informational perspective are rich enough to encode all information on the past which is future-relevant in one (present) state. By Treur (2007a, b), by many examples from Physics or from Cognitive Science it is illustrated how temporal factorisation works for world dynamics. Note: the converse implication

$$\exists p \quad [a \Rightarrow p \quad \& \quad p \Rightarrow b] \Rightarrow [a \Rightarrow b] \tag{16.5}$$

of temporal factorisation is a logically valid statement in an almost trivial manner (by transitivity of implication), and as such does not add new information. In contrast, the temporal factorisation principle itself is nontrivial. Temporal factorisation can also be applied iteratively so that also $a \Rightarrow p$ and $p \Rightarrow b$ are factorised further in

$$a \Rightarrow p_1 \quad \& \quad p_1 \Rightarrow p \qquad p \Rightarrow p_2 \quad \& \quad p_2 \Rightarrow b \tag{16.6}$$

In this way, temporal factorisation basically expresses that to describe in more detail global flows of dynamics from patterns in the past to patterns in the future, intermediate states can be found that in a sense break up the global process in smaller steps. These intermediate states carry informational content about the past and about the future.

An interesting question is how temporal factorisation relates to perspectives on dynamics from Descartes (1634), Laplace (1825), Ashby (1952), van Gelder and Port (1995). This will be discussed in some detail. From these, Descartes (1634) claims that there are 'laws of nature' as relationships between world states for different points in time, in the form that past world states determine future world states. He emphasizes that only the laws of nature determine the dynamics of the world (after a starting time):

> Know, then, first that by "nature" I do not here mean some deity or other sort of imaginary power. Rather, I use that word to signify matter itself, insofar as I consider it taken together with all the qualities that I have attributed to it, and under the condition that God continues to preserve it in the same way that He created it. For from that alone (i.e., that He continues thus to preserve it) it follows of necessity that there may be many changes in its parts that cannot, it seems to me, be properly attributed to the action of God (because that action does not change) and hence are to be attributed to nature. The rules according to which these changes take place I call the "laws of nature". (Descartes 1634, Chap. 7: On the Laws of Nature of this New World)

This is also called the clockwork universe view. It describes how relationships between world states over time (laws of nature) drive world dynamics, in the sense that patterns for past world states a determine patterns for future world states b, or $a \Rightarrow b$.

Laplace (1825)'s view is slightly different and a bit more refined; it is summarized by this quote:

> We may regard the present state of the universe as the effect of its past and the cause of its future. (Laplace 1825)

Both views of Descartes and Laplace describe a deterministic world. Although temporal factorisation somehow relates to the above two perspectives, in contrast it does not claim a fully deterministic world. That principle just applies to world aspects that are deterministic; this is indicated by the condition $a \Rightarrow b$.

To make the relations to temporal factorisation more precise, Descartes' view is formulated by

$$\begin{aligned} &\text{D: Dynamics works by temporal past - to - future} \\ &\quad \text{relationships } a \Rightarrow b. \end{aligned} \tag{16.7}$$

and Laplace's view is formulated by

$$\begin{aligned} &\text{L: Dynamics works by temporal past - to - present} \\ &\quad \text{and present - to - future relationships } a \Rightarrow p \quad \text{and} \quad p \Rightarrow b. \end{aligned} \tag{16.8}$$

So, D relates to the antecedent of the principle of Temporal Factorisation TF, and L to the consequent; temporal factorisation logically connects D and L in the following way: Descartes' view D and TF together (by Modus Ponens) entail Laplace's L:

$$\begin{aligned} &\text{D}: \quad a \Rightarrow b \\ &\text{TF}: \quad [a \Rightarrow b] \Rightarrow \exists p [a \Rightarrow p \quad \& \quad p \Rightarrow b] \\ &\text{- - - - - - - - - - - Modus Ponens - - - - - - - - - - -} \\ &\text{L}: \quad \exists p \ [a \Rightarrow p \ \ \& \ \ p \Rightarrow b] \end{aligned} \tag{16.9}$$

which may be summarized as

$$\text{D} \ \& \ \text{TF} \Rightarrow \text{L} \tag{16.10}$$

is logically true. It turns out that temporal factorisation is a silent assumption that explains the shift from Descartes's perspective D to Laplace's perspective L. While these two perspectives both assume deterministic world processes, temporal factorisation itself does not, due to the conditional. It therefore is more general and also applicable for nondeterministic (parts of the) world(s) for which D and L do not hold.

In (van Gelder and Port, 1995), after Ashby (1952) dynamics is based on the concept of state-determined system:

> … its current state always determines a unique future behaviour … the future behaviour cannot depend in any way on whatever states the system might have been in *before* the current state. (van Gelder and Port 1995), p. 6

This can be described by the implication $p \Rightarrow b$, which focuses on the step from present state p to future pattern b, while it does not consider how any past pattern a leads to the present state p.

One example, in a more or less similar form also used in (Treur 2007a), illustrating temporal factorisation is the following. Imagine in a real-world or gaming situation in the current state a door is encountered that was locked by somebody in the past and not unlocked since then. It then can be opened in the future when someone brings the right key. So, the past pattern a can be described as

At some past time point someone locked the door and since then nobody unlocked it (16.11)

and the future pattern b as

If at some future time point someone carries the right key,
then the door can be unlocked (16.12)

Moreover, mediating state property p is here

The door is locked (16.13)

From the temporal relationship $a \Rightarrow b$, by temporal factorisation it can be derived that some mediating state property p exists for which the temporal past-to-present relationship $a \Rightarrow p$ and the present-to-future relationship $p \Rightarrow b$ hold. In this case, this state property p specifies that the door is locked by its specific lock. Here $a \Rightarrow p$ specifies that if somebody locked the door in the past and nobody unlocked it since then, it is locked in the present, while $p \Rightarrow b$ expresses that if it is locked in the present, when someone brings the appropriate key in the future it can be unlocked.

Viewed informationally, within the world in the present state p, the lock locking the door represents information about the lock, and when in the future the appropriate key in accordance with that lock information, is brought, the door unlocks. This case illustrates how the mediating state property p encodes information, and this information determines the world's future by pattern b, described as 'when the right key is brought, the door unlocks'. Here humans are the actors whose actions encode the information in the physical world, since the lock and also the key were made by humans. So, in this case humans informationalise the world, or, in other words, the world becomes more informational because of human intervention. Another example of such human-intervened informationalisation of the world can be found in the story

of Little Thumb who drops pebbles to mark his route so that he can find back his way home. In contrast to such human-intervened cases, according to the temporal factorisation principle the world (as an actor itself) is creating and exploiting a similar encoding in present world state configurations of information on past patterns without human intervention.

16.2.2 Criterial Causation

Neural processing involves propagation of activation of neurons; e.g., (Hebb 1949; Tse 2013). This propagation depends on network characteristics of the network of neurons. For example, it depends on synaptic connection strengths for connections between neurons and on the excitability thresholds for neurons. The configuration in the brain defined by these characteristics serves as criteria for the signals, that are incoming for a neuron, to make it fire, which by Tse (2013, 2018) is termed *criterial causation*. Here the criteria represent *information* represented in the specific brain configuration: 'physically realised informational criteria' (Tse 2013, p. 259); in the future, the neuron will fire when these criteria are fulfilled.

> I have proposed a three-stage model of a neuronal mechanism that underlies mental causation and free will, according to which (1) new physical/informational criteria are set in a neuronal circuit on the basis of preceding physical/mental processing at t_1, in part via a mechanism of rapid resetting that effectively changes the inputs to a postsynaptic neuron. These changes can be driven volitionally or nonvolitionally, depending on the neural circuitry involved. (2) At t_2, inherently variable inputs arrive at the postsynaptic neuron, and (3) at t_3 physical/informational criteria are met or not met, leading to postsynaptic neural firing or not. (Tse 2013, p. 14; pp. 148–149)

So, neural dynamics, does not concern only the activation levels of neurons, it also concerns how network characteristics representing the above-mentioned criteria change over time: both dynamics *within* the network and dynamics *of* the network occur. Changes of network characteristics are created by patterns in the past affecting them. These changes might be slow in comparison to activation propagation of neurons, as in learning from of multiple experiences, but might also be very fast, almost instantly, as in creation of memories: *rapid resetting* of the criteria (Tse 2013). This form of network adaptation describes *emerging information*; the activation of neurons takes place based on this obtained informational content (Tse 2013, 2018).

> [In addition to activation propagation] … neurons can also rapidly and dynamically change the weights, gains, and temporal integration properties of synapses of other neurons without necessarily triggering an action potentional in those neurons (§4.54-4.60). This and other physical mechanisms accomplish a recoding of the inputs that will make a neuron fire in the future. Such recoding changes both the physical and the informational criteria that a neuron places on its inputs, even when the threshold for firing at the axon hillock remains constant. (Tse 2013, p. 22)

He emphasizes that patterns over time have a causal effect, not just states. The criteria play an important role in decoding these patterns:

> Patterns in input can be genuinely causal only if there are physical detectors, such as neurons, that respond to patterns in input and then change the physical system in which they reside if the criteria for the presence of a pattern have been met. (Tse 2013, p. 9)
>
> Criterial decoders allow for the emergence of physically realized informational causal chains that are not just physical causal chains (although they are that), but also, in the case of the brain, mental causal chains. What makes one outcome occur rather than others is that certain informational criteria were met in decoders that then triggered consequences that would have differed had these informational criteria not been met. (Tse 2013, p. 116)

During evolution these mechanisms have been developed not in an arbitrary manner but because of their informational content:

> The whole point of setting up a physical causal system to react criterially to input, as we have in the brain, is to make sure that only the subset of possible physical causal chains that are also informational chains are the ones that will be realized. Far from being an incidental fact about this subset of physical causal chains, natural selection has operated to optimize the efficiency of information processing within possible informational causal chains to fit the needs of an organism within its niche. The fact that information processing is realized in physical causal chains is, in a sense, beside the point. An animal does not get weeded out because it followed such and such a physical causal chain as opposed to another; it gets weeded out because it failed to recognize the lion where there was one. It gets weeded out or survives to the extent that it processes information poorly or well. Evolution of course operates not only to weed out suboptimal bodies, but also suboptimal perceptual, emotional, and cognitive systems realized in neuronal and endocrine system activity. (Tse 2013, p. 131)

16.2.3 How Criterial Causation Relates to Temporal Factorisation

From Sect. 16.1 and the explanations above, it is clear that the notions of temporal factorisation and of criterial causation are in a close relationship; also the adopted pictures in Figs. 16.1 and 16.2 for these notions show similarity. The correspondence is as follows, using the above illustration for temporal factorisation by the lock and key example:

(1) In past pattern *a* someone locked the door, and since then nobody unlocked the door
(2) This creates in the present state mediating state property *p* corresponding to the locked door with a specific lock
(3) When in a future pattern *b* someone attempts to unlock the door with the right key
(4) When it indeed fits in future pattern *b*, the door unlocks.

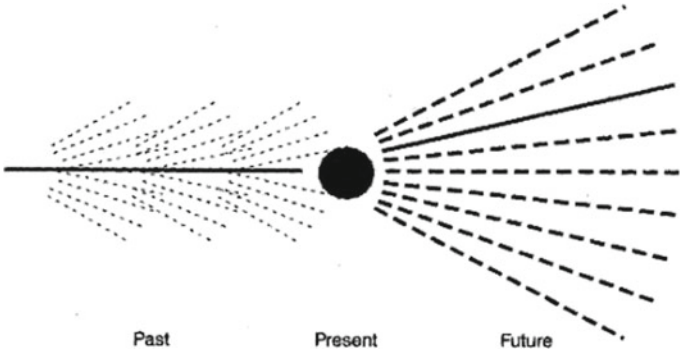


Fig. 16.2 Criterial causation; picture adopted from Tse (2013, p. 125, Fig. 2)

This corresponds to criterial causation as follows

(1) In past pattern a, someone locking the door (and nobody unlocking it in the meantime) corresponds to setting the criteria for criterial causation
(2) The mediating state property p corresponds to the criteria set
(3) In a future pattern b, attempting to fit a key of the right shape in the lock, corresponds to checking of the fulfillment of the criteria.
(4) If after a fit of key and lock in future pattern b (= the criteria are indeed fulfilled), the door unlocks; this corresponds to activation of the neuron in pattern b.

More specifically, criterial causation as described in Tse (2013) covers the consequent

$$\exists p \;\; [a \Rightarrow p \;\; \& \;\; p \Rightarrow b] \tag{16.14}$$

of the principle of temporal factorisation, without mentioning a condition $a \Rightarrow b$, similar to Laplace's description L (16.8). So, if criterial causation is indicated by CC, then the following logical relationship is obtained:

$$\mathrm{D} \;\; \& \;\; \mathrm{TF} \Rightarrow \mathrm{CC} \tag{16.15}$$

But note that criterial causation specifically addresses dynamics within the brain and not of the world in general as TF does, and also D (16.7) and L (16.8) do. Moreover, note that Tse (2013) also mentions that his perspective might cover small random fluctuations as well. However, for now this is left out of the analysis here.

16.3 Network-Oriented Modeling for Adaptive Networks

Both temporal factorisation and Tse (2013)'s view concerning criterial causation address dynamics and adaptation. A specific modeling approach addressing dynamics and adaptivity is the network-oriented modeling approach described in Treur (2016, 2020b). The current section focuses on briefly describing this general network modeling approach. In Sects. 16.4–16.6 this modeling approach is applied and related to temporal factorisation and criterial causation.

16.3.1 Network Models

According to the network-oriented modeling approach described in (Treur 2016,2020b) a network model is characterised by:

- **connectivity characteristics**

  Connections from a node (or state) X to a node Y and their *weights* $\boldsymbol{\omega}_{X,Y}$

- **aggregation characteristics**
  For any node Y, some *combination function* $\mathbf{c}_Y(..)$ defines aggregation that is applied to the single impacts $\boldsymbol{\omega}_{X,Y}X(t)$ on Y through its incoming connections from states X

- **timing characteristics**
  Each node Y has a *speed factor* $\boldsymbol{\eta}_Y$ defining how fast it changes for given (aggregated) impact

The difference (or differential) equations that are useful for simulation purposes and also for analysis of network dynamics incorporate these network characteristics $\boldsymbol{\omega}_{X,Y}$, $\mathbf{c}_Y(..)$, $\boldsymbol{\eta}_Y$: it holds

$$Y(t+\Delta t) = Y(t) + \boldsymbol{\eta}_Y\left[\mathbf{c}_Y\left(\boldsymbol{\omega}_{X_1,Y}X_1(t), \ldots, \boldsymbol{\omega}_{X_k,Y}X_k(t)\right) - Y(t)\right]\Delta t \quad (16.16)$$

for any state Y and where $X_1, \ldots, X_k$ are the states from which it gets its incoming connections. Some examples of useful combination functions are:

- the advanced logistic sum function $\mathbf{alogistic}_{\sigma,\tau}(..)$ defined by:

$$\mathbf{alogistic}_{\sigma,\tau}(V_1, \ldots, V_k) = \left[\frac{1}{1+\mathrm{e}^{-\sigma(V_1+\cdots+V_k-\tau)}} - \frac{1}{1+\mathrm{e}^{\sigma\,\tau}})\right](1+\mathrm{e}^{-\sigma\,\tau}) \quad (16.17)$$

- the simple logistic sum function $\mathbf{alogistic}_{\sigma,\tau}(..)$ defined by:

$$\mathbf{slogistic}_{\sigma,\tau}(V_1, \ldots, V_k) = \frac{1}{1 + e^{-\sigma(V_1 + \cdots + V_k - \tau)}} \tag{16.18}$$

The above concepts enable to design network models and their dynamics in a declarative manner, based on mathematically defined functions and relations.

16.3.2 Modeling Adaptive Networks as Self-Modeling Networks

Realistic network models are usually adaptive: their network characteristics often are adapted over time. Therefore, their dynamics is usually an interaction (sometimes called co-evolution) of these two sorts of dynamics: dynamics of the nodes (or states) in the network (dynamics *within* the network) versus dynamics of the characteristics of the network (dynamics *of* the network). Dynamics of the network's nodes are modeled declaratively by declarative mathematical functions and relations. In contrast, the dynamics of the network characteristics traditionally are described in a procedural, algorithmic nondeclarative manner, which then leads to a hybrid type of model. But by using *self-models* within the network, a network-oriented conceptualisation can also be applied to *adaptive* networks to obtain a declarative description using mathematically defined functions and relations; see (Treur 2020b). This works through the addition of new nodes to the network (called self-model states or reification states) which represent (adaptive) network characteristics. Such nodes are depicted at a next level (*self-model level*), where the original network is at a *base level*. These types of characteristics with their self-model states and their roles are shown in Table 16.1.

Table 16.1 Different network characteristics and self-model states for them

Types of characteristics	Concepts	Notations	Self-model states	Role played by the self-model state
Connectivity characteristics	Connections weights	$\boldsymbol{\omega}_{X,Y}$	$\mathbf{W}_{X,Y}$	Connection weight **W**
Aggregation characteristics	Combination functions and their parameters	$\mathbf{c}_Y(..)$ $\boldsymbol{\pi}_{i,j,Y}$	$\mathbf{C}_{i,Y}$ $\mathbf{P}_{i,j,Y}$	Combination function weight **C** Combination function parameter **P**
Timing characteristics	Speed factors	$\boldsymbol{\eta}_Y$	$\mathbf{H}_Y$	Speed factor **H**

This provides an extended network, also called *self-modeling network*. Like for all network models, a self-modeling network model is specified in a (network-oriented) declarative mathematical manner based on nodes and connections. These include interlevel connections relating nodes at one level to nodes on the other.

The outcome is also a network model (Treur 2020a, Chap. 10). This whole construction can be applied iteratively to obtain multiple self-model levels that can provide higher-order adaptive networks, and is quite useful to model, for example, plasticity and metaplasticity in the form of a second-order adaptive network with three levels, one base level and a first- and a second-order self-model level; e.g., (Abraham and Bear 1996; Treur 2020b) or (Treur 2020a), Chap. 4.

To support the design of network models, for any application from a library predefined basic combination functions $\mathrm{bcf}_i(..)$, $i = 1,..,m$ are selected by assigning weights $\boldsymbol{\gamma}_{i,Y}$, where the combination function then becomes the weighted average

$$\mathbf{c}_Y(..) = \left(\boldsymbol{\gamma}_{1,Y}\mathrm{bcf}_1(..) + \ldots + \boldsymbol{\gamma}_{m,Y}\mathrm{bcf}_m(..)\right)/\left(\boldsymbol{\gamma}_{1,Y} + \ldots + \boldsymbol{\gamma}_{m,Y}\right) \qquad (16.19)$$

Furthermore, parameters of combination functions are specified, so that $\mathrm{bcf}_i(..) = \mathrm{bcf}_i(\mathbf{p},\mathbf{v})$ where $\mathbf{p}$ is a list of parameters and $\mathbf{v}$ is a list of values.

16.4 Temporal Factorisation Modeled by Networks

As noted in Sect. 16.2.1, temporal factorisation basically expresses that to describe in more detail global flows of dynamics from patterns in the past to patterns in the future, intermediate states can be found that in a sense break up the global process in smaller steps. This can be related to networks of states that together describe the overall process. How this works, will be shown in the current section for two examples that are used to illustrate the temporal factorisation principle.

16.4.1 An Example Network Model Illustrating Temporal Factorisation

The first example scenario used here to relate temporal factorisation to network models is the locked door example described in the last paragraph of Sect. 16.2.1. It was modeled by the small network with basic connectivity as shown in Fig. 16.3. The states can be given the following meaning as shown in Table 16.2, second column. A second example about animal behaviour will be explained below in the last two paragraphs of the current section.

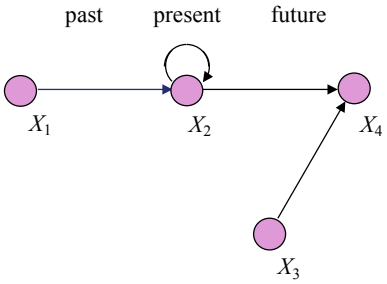


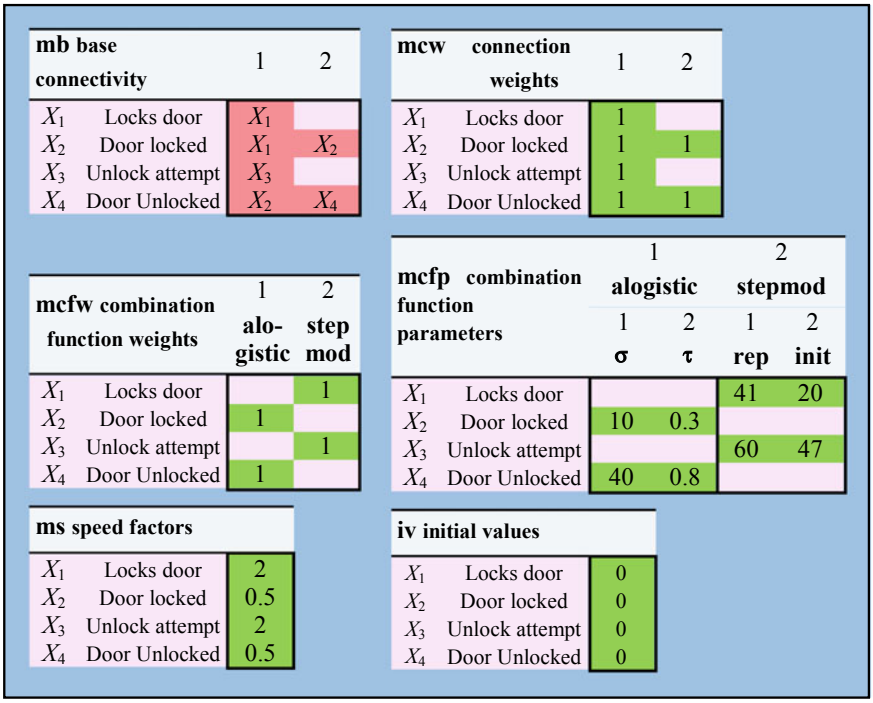
Fig. 16.3 Network model illustrating the two temporal factorisation examples

Table 16.2 Explanation of the states for the two examples

State		Explanation locked door example	Animal example	
nr	Name			
X_1	Locks door	Someone locks the door	Mouse sees food	Mouse sees food at location l_1
X_2	Door locked	Locked door	Mental state for food	Mouse has mental state for food at l_1
X_3	Unlock attempt	Attempt to unlock the door by the specific key	Mouse released	Mouse is released and free to go
X_4	Door unlocked	Unlocked door	Mouse goes to l_1	Mouse goes to l_1

In this network, X_2 is the state (locked door, with a specific lock) that is considered in the present. There is one pathway from the past, coming from X_1, (someone locks the door) and a pathway for the future involving X_3 (someone attempts to unlock the door, with a specific key) and X_4 (unlocked door). The circular arrow for X_2 models persistence of that state (locked door), assures that nobody was unlocking the door since it was locked.

In Fig. 16.4 the specification in role matrix format is shown. Any role matrix has rows for all the states of the network. Within each row the impacts on that state from that role are listed. For example, in role matrix **mb** addressing *base connectivity* in the row for X_4 it is indicated that it gets impact from states X_2 and X_3. Connectivity role matrix **mcw** addressing connection weights indicates, for example, that X_4 gets impact from connection weights 1 from X_2 and X_3, respectively. *Aggregation* role matrix **mcfw** specifying combination function weights indicates that multiple impacts are aggregated by the advanced logistic sum function $\mathbf{alogistic}_{\sigma,\tau}(..)$ defined by (16.17). A second combination function is **stepmod**, which is used to get independent events happen regularly (at time **init**) according to a cyclic pattern with a certain repetitive duration (duration **rep**). These two combination functions are numbers 2 and 35 of the Combination Function Library; this selection from the library is indicated by **mcf** = [2 35], after which these combination functions have numbers 1



mb base connectivity

		1	2
X_1	Locks door	X_1	
X_2	Door locked	X_1	X_2
X_3	Unlock attempt	X_3	
X_4	Door Unlocked	X_2	X_4

mcw connection weights

		1	2
X_1	Locks door	1	
X_2	Door locked	1	1
X_3	Unlock attempt	1	
X_4	Door Unlocked	1	1

mcfw combination function weights

		1 alo-gistic	2 step mod
X_1	Locks door		1
X_2	Door locked	1	
X_3	Unlock attempt		1
X_4	Door Unlocked	1	

mcfp combination function parameters

		1 alogistic		2 stepmod	
		1 σ	2 τ	1 rep	2 init
X_1	Locks door			41	20
X_2	Door locked	10	0.3		
X_3	Unlock attempt			60	47
X_4	Door Unlocked	40	0.8		

ms speed factors

X_1	Locks door	2
X_2	Door locked	0.5
X_3	Unlock attempt	2
X_4	Door Unlocked	0.5

iv initial values

X_1	Locks door	0
X_2	Door locked	0
X_3	Unlock attempt	0
X_4	Door Unlocked	0

Fig. 16.4 Role matrices specification and initial values for the network model with connectivity depicted in Fig. 16.3 illustrating temporal factorisation

and 2 for this model. In *aggregation* role matrix **mcfp** for the combination function parameters it is indicated that the parameter values for this function are 40 (for steepness **σ**) and 1.8 (for threshold **τ**), respectively. Finally, in the role matrix **ms** for *timing* it is indicated that X_4 has speed factor 0.5, and in **iv** it is indicated that all initial values are 0.

16.4.2 Simulation for the Network Model Illustrating Temporal Factorisation

In Fig. 16.5 a simulation for this network model is shown. The present is some time point between 41 and 45. As can be seen, someone locks the door between time 20 and time 40 (blue line), which is considered the past here. As a consequence the door is locked (red line) by a specific type of lock, and this state persists (nobody unlocks it until the present). Therefore this locked door is also there in the present. In the future, from time 47 to 59 someone brings the right key after time 47 and attempts

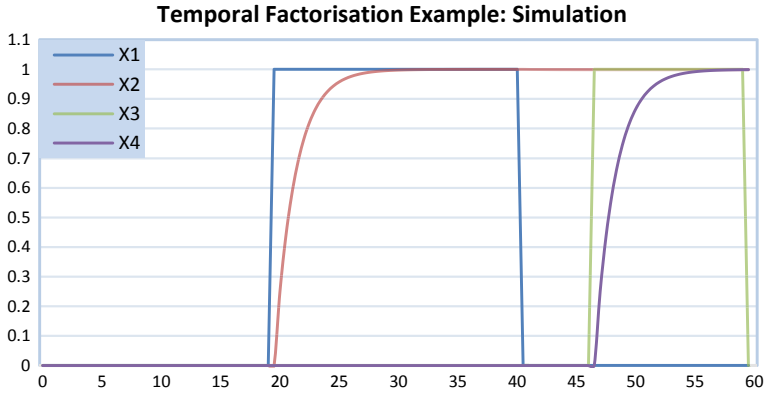


Fig. 16.5 Simulation for the network model for temporal factorisation of Fig. 16.3

to unlock the door (green line), given the specific lock fitting to the key (represented by X_3), resulting in an unlocked door (purple line).

16.4.3 Application of the Network Model to Delayed Response Behaviour

Another illustration of temporal factorisation uses the notion of 'delayed response behaviour' of animals (often considered for mice), also used in (Treur 2007); e.g., (Hunter 1912; Tinklepaugh 1932). Consider the following concepts for an experiment in which the animal is first kept behind a window and later on released:

c food at location l_0 is visible for a given animal
d the animal is released
e the animal is at l_0.

In such experiments, after having seen the food, the food is covered by a cup so that it is not visible anymore. A past pattern *a* can be described by:

For at least one time point in the past, state *c*
(denoting visible food at l_0) occurred and since then
it was not visible that it did not occur anymore. (16.20)

A future pattern *b* can be described by:

If in future the state *d* occurs (denoting that animal is released),
then at a later time the state *e* will occur (denoting that animal is at l_0). (16.21)

From various experiments it turns out that always when past trace a occurs, also future trace b occurs: $a \Rightarrow b$. Therefore, temporal factorisation applies to these patterns: since $a \Rightarrow b$ holds, we know that a state property p exists for which the temporal relationships $a \Rightarrow p$ and $p \Rightarrow b$ hold. This postulated state property p is a cognitive state, which functions as a form of memory. In other words, after observing in many animal experiments past-future relationship $a \Rightarrow b$, the principle of temporal factorisation claims that some mental (memory) state emerges in the present after past pattern a. This mental state in the present determines the animal's future behaviour so that b holds. So, the mediating state property p can be formulated for this case as:

$$\text{The animal has a mental (memory) state for food at } l_0 \tag{16.22}$$

Such a memory state encodes information about the animal's environment, and this information determines the future pattern. For this case the information formation is a human-independent emerging process which takes place without human intervention; even for the mouse it is an automatic process and not intentional. Therefore in this case the world itself is the actor that does information formation, through the brain's mechanisms.

To model this example, the same network as above can be used, but this time the states are interpreted as follows:

Mouse sees food at l_0	X_1
Mouse′s mental state with food at l_0	X_2
Mouse is released	X_3
Mouse goes to l_0	X_4

So, now the simulation story for Fig. 16.4 is as follows. The mouse sees food at l_0 in the past, between time 20 and time 40 (blue line). Due to this, it forms a mental (memory) state for the food at l_0 (red line). As it was not seen that the food disappeared, this mental state persists and is also there in the present, at some time point between 41 and 45. In the future, from time 47 to 59 the mouse is released after time 47 (green line) and due to this and the mentioned mental state, it goes to l_0 (purple line).

16.5 Modeling Criteria for Criterial Causation for Network Models

In this section a first step is made to model criterial causation by network models: the criteria will be analysed in more detail. Based on difference Eq. (16.16), the activation criterion for a state Y can be formulated by requiring that the aggregated

impact $\mathbf{c}_Y(\boldsymbol{\omega}_{X_1,Y}X_1(t), \ldots, \boldsymbol{\omega}_{X_k,Y}X_k(t))$ on Y has a certain level at least higher than 0.5:

$$\mathbf{c}_Y(\boldsymbol{\omega}_{X_1,Y}X_1(t), \ldots, \boldsymbol{\omega}_{X_k,Y}X_k(t)) > 0.5 \tag{16.23}$$

This (16.23) is used here as the general criterion in a network model for criterial causation for any arbitrary combination function $\mathbf{c}_Y(..)$ for any state Y.

16.5.1 Criteria Using Logistic Combination Functions

Examples of combination functions $\mathbf{c}_Y(...)$ are the simple logistic $\mathbf{slogistic}_{\sigma,\tau}(\ldots)$ and advanced logistic function $\mathbf{alogistic}_{\sigma,\tau}(\ldots)$, each with steepness parameter $\boldsymbol{\sigma} > 0$ and excitability threshold parameter $\boldsymbol{\tau}$; see (16.17), (16.18). For each of the combination functions criterion (16.23) can be made more specific: for example, for $\mathbf{slogistic}_{\sigma,\tau}(\ldots)$ (16.18), criterion (16.23) can be reformulated into the more specific criterion

$$\frac{1}{1+\mathrm{e}^{-\boldsymbol{\sigma}(V_1+\cdots+V_k-\boldsymbol{\tau})}} > 0.5 \tag{16.24}$$

with the V_i the single impacts $\boldsymbol{\omega}_{X_i,Y}X_i(t)$ on state Y. By algebraic rewriting (shown in Box 16.1, left), this translates into the following linear inequality in $X_1(t), \ldots, X_k(t)$:

$$\boldsymbol{\omega}_{X_1,Y}X_1(t) + \cdots + \boldsymbol{\omega}_{X_k,Y}X_k(t) > \boldsymbol{\tau} \tag{16.25}$$

So, (16.25) is the criterion in particular for $\mathbf{slogistic}_{\sigma,\tau}(\ldots)$**.** Similarly, for $\mathbf{alogistic}_{\sigma,\tau}(\ldots)$, see (16.17), the following criterion is found (see Box 16.1, right), which still is a linear inequality in $X_1(t), \ldots, X_k(t)$:

$$\boldsymbol{\omega}_{X_1,Y}X_1(t) + \cdots + \boldsymbol{\omega}_{X_k,Y}X_k(t) > \mathrm{c} \tag{16.26}$$

$$\text{with } \mathrm{c} = \tau - \log\left(\frac{1}{\frac{0.5}{1+\mathrm{e}^{-\sigma\tau}} + \frac{1}{1+\mathrm{e}^{\sigma\tau}}} - 1\right)/\sigma$$

It can be seen that for these two specific combination functions, criteria (16.25) and (16.26) are linear inequalities concerning state values $X_1, \ldots, X_k$ with as constants expressions in terms of network characteristics $\boldsymbol{\omega}$, $\boldsymbol{\sigma}$, $\boldsymbol{\tau}$.

Box 16.1 Deriving the criteria (16.25) (left) and (16.26) (right) for criterial causation in case of combination function $\mathbf{slogistic}_{\sigma,\tau}(\ldots)$ or $\mathbf{alogistic}_{\sigma,\tau}(\ldots)$.

$1 + e^{-\sigma(V_1+\cdots+V_k-\tau)} < 2 \Leftrightarrow$
$e^{-\sigma(V_1+\cdots+V_k-\tau)} < 1 \Leftrightarrow$
$\sigma(V_1 + \cdots + V_k - \tau) > 0 \Leftrightarrow$
$V_1 + \cdots + V_k > \tau \Leftrightarrow$
$\omega_{X_1,Y} X_1(t) + \cdots + \omega_{X_k,Y} X_k(t) > \tau$

$\left[\frac{1}{1+e^{-\sigma(V_1+\cdots+V_k-\tau)}} - \frac{1}{1+e^{\sigma\tau}}\right](1 + e^{-\sigma\tau}) > 0.5 \Leftrightarrow$

$\left[\frac{1}{1+e^{-\sigma(V_1+\cdots+V_k-\tau)}} - \frac{1}{1+e^{\sigma\tau}}\right] > \frac{0.5}{1+e^{-\sigma\tau}} \Leftrightarrow$

$\frac{1}{1+e^{-\sigma(V_1+\cdots+V_k-\tau)}} > \frac{0.5}{1+e^{-\sigma\tau}} + \frac{1}{1+e^{\sigma\tau}} \Leftrightarrow$

$1 + e^{-\sigma(V_1+\cdots+V_k-\tau)} < \frac{1}{\frac{0.5}{1+e^{-\sigma\tau}} + \frac{1}{1+e^{\sigma\tau}}} \Leftrightarrow$

$e^{-\sigma(V_1+\cdots+V_k-\tau)} < \frac{1}{\frac{0.5}{1+e^{-\sigma\tau}} + \frac{1}{1+e^{\sigma\tau}}} - 1 \Leftrightarrow$

$-\sigma(V_1 + \cdots + V_k - \tau) < \log\left(\frac{1}{\frac{0.5}{1+e^{-\sigma\tau}} + \frac{1}{1+e^{\sigma\tau}}} - 1\right)$
$\Leftrightarrow$
$\sigma(V_1 + \cdots + V_k - \tau) > -\log\left(\frac{1}{\frac{0.5}{1+e^{-\sigma\tau}} + \frac{1}{1+e^{\sigma\tau}}} - 1\right)$
$\Leftrightarrow$
$V_1 + \cdots + V_k > \tau - \log\left(\frac{1}{\frac{0.5}{1+e^{-\sigma\tau}} + \frac{1}{1+e^{\sigma\tau}}} - 1\right) \Leftrightarrow$

$\omega_{X_1,Y} X_1(t) + \cdots + \omega_{X_k,Y} X_k(t) >$
$\tau - \log\left(\frac{1}{\frac{0.5}{1+e^{-\sigma\tau}} + \frac{1}{1+e^{\sigma\tau}}} - 1\right)/\sigma$

16.5.2 Criteria Using Other Combination Functions

In a similar way, the criteria for other combination functions were found: scaled minimum $\mathbf{smin}_{\lambda}(..)$ and scaled maximum $\mathbf{smax}_{\lambda}(..)$, and scaled sum $\mathbf{ssum}_{\lambda}(..)$, Euclidean $\mathbf{eucl}_{n,\lambda}(..)$ and scaled geometric mean $\mathbf{sgeomean}_{\lambda}(..)$; see Table 16.3. Note that all of the criteria, except the last two are linear inequalities, whereas the last two are inequalities involving powers and products of $X_1(t), \ldots, X_k(t)$.

As real-world network models are often adaptive, the network characteristics mentioned can and usually do change over time. In this manner, the criteria and the informational content represented by the criteria emerge dynamically. For example, in criteria (16.23), (16.25), (16.26) the connection weights $\boldsymbol{\omega}$, the excitability threshold $\boldsymbol{\tau}$ and the steepness $\boldsymbol{\sigma}$ may change over time, by which they change the criterion. This

Table 16.3 Overview of criteria for criterial causation for a few combination functions

Combination function name and formula		Criterion for criterial causation
Name	$\mathbf{c}_Y(V_1, \ldots, V_k)$	$\mathbf{c}_Y(\boldsymbol{\omega}_{X_1,Y} X_1(t), \ldots, \boldsymbol{\omega}_{X_k,Y} X_k(t)) > 0.5$
slogistic$_{\sigma,\tau}(V_1, \ldots, V_k)$	$\frac{1}{1+e^{-\sigma(V_1+\cdots+V_k-\tau)}}$	$\boldsymbol{\omega}_{X_1,Y} X_1(t) + \cdots + \boldsymbol{\omega}_{X_k,Y} X_k(t) > \boldsymbol{\tau}$
alogistic$_{\sigma,\tau}(V_1, \ldots, V_k)$	$\left[\frac{1}{1+e^{-\sigma(V_1+\cdots+V_k-\tau)}} - \frac{1}{1+e^{\sigma\tau}}\right](1 + e^{-\sigma\tau})$	$\boldsymbol{\omega}_{X_1,Y} X_1(t) + \cdots + \boldsymbol{\omega}_{X_k,Y} X_k(t) > \boldsymbol{\tau} - \log\left(\frac{1}{\frac{0.5}{1+e^{-\sigma\,\boldsymbol{\tau}}} + \frac{1}{1+e^{\sigma\,\boldsymbol{\tau}}}} - 1\right)/\sigma$
smax$_{\boldsymbol{\lambda}}(V_1, \ldots, V_k)$	$\mathbf{max}(V_1, \ldots, V_k)\,/\boldsymbol{\lambda}$	$\boldsymbol{\omega}_{X_i,Y} X_i(t) > 0.5\ \boldsymbol{\lambda}$ for some i
smin$_{\boldsymbol{\lambda}}(V_1, \ldots, V_k)$	$\mathbf{min}(V_1, \ldots, V_k)\,/\boldsymbol{\lambda}$	$\boldsymbol{\omega}_{X_i,Y} X_i(t) > 0.5\ \boldsymbol{\lambda}$ for all i
ssum$_{\boldsymbol{\lambda}}(V_1, \ldots, V_k)$	$\frac{V_1+\cdots+V_k}{\boldsymbol{\lambda}}$	$\boldsymbol{\omega}_{X_1,Y} X_1(t) + \cdots + \boldsymbol{\omega}_{X_k,Y} X_k(t) > 0.5\boldsymbol{\lambda}$
eucl$_{\boldsymbol{n},\boldsymbol{\lambda}}(V_1, \ldots, V_k)$	$\sqrt[n]{\frac{V_1^n+\cdots+V_k^n}{\boldsymbol{\lambda}}}$	$\boldsymbol{\omega}^{\boldsymbol{n}}_{X_1,Y} X_1(t)^{\boldsymbol{n}} + \cdots + \boldsymbol{\omega}^{\boldsymbol{n}}_{X_k,Y} X_k(t)^{\boldsymbol{n}} > 0.5^{\boldsymbol{n}}\ \boldsymbol{\lambda}$
sgeomean$_{\lambda}(V_1, \ldots, V_k)$	$\sqrt[k]{\frac{V_1 * \ldots * V_k}{\boldsymbol{\lambda}}}$	$\boldsymbol{\omega}_{X_1,Y} X_1(t) * \ldots * \boldsymbol{\omega}_{X_k,Y} X_k(t) > 0.5^{\boldsymbol{k}}\ \boldsymbol{\lambda}$

can be modeled well in the form of an adaptive network. Using for this a self-modeling network, such adaptivity needed for criterial causation can easily be modeled. In that setting, for temporal factorisation a mediating state property p is specified by the values of some states, which can be base level states as in Sect. 16.4, or states at any self-model level, as will be illustrated in Sects. 16.6 and 16.7. Also the past patterns a and the future patterns b can be specified by state value assigments for a number of states, in this case over time. In relation to criteria (16.25) and (16.26), the connection weight coefficients and threshold constants are represented by self-model states, whereas the states $X_1, \ldots, X_k$ in the criteria are base level states.

16.6 How a Developing Mental Model Creates Emergent Information in the Brain

As explained in Sect. 16.3, a network model is defined by three main types of network structure characteristics (connectivity, aggregation, timing), which are modeled by $\boldsymbol{\omega}_{X,Y}$, $\mathbf{c}_Y(..)$, $\boldsymbol{\eta}_Y$. The canonical difference equation as expressed in (16.16) above includes these network characteristics. For an adaptive network some of these characteristics are explicitly represented by self-model states $\mathbf{W}_{X,Y}$, $\mathbf{C}_Y$, $\mathbf{P}_Y$ or $\mathbf{H}_Y$ and used in this difference equation accordingly, so that they can become adaptive.

16.6.1 An Example Scenario for Learning and Use of a Mental Model

The scenario used for this section concerns learning of a mental model by a new person in a company who has to learn to recognize colleagues. It goes as follows.

Example Scenario

A new person in a company has to learn to recognize a colleague from only seeing his face; this face is stimulus *s*. Two colleagues a_1 and a_2 are assumed that are options to choose from. Picking one of them is indicated by activation of sensory representation state srs_{a_i}. A belief bs_1 suggests that it is colleague a_1, and a belief bs_2 that it is colleague a_2. These beliefs are only meant indicative (e.g., based on the location at which the person is encountered), but not sufficient to decide for one of them. As the beliefs and *s* are triggered by independent circumstantial factors, for the network model they just happen. Two types of network characteristics are addressed as adaptive: the weights of the connections from sensory representation srs_s for *s* to srs_{a_1} and srs_{a_2}, and the excitability thresholds for states srs_{a_1} and srs_{a_2}. The type of adaptation of associations between sensory representations relates to the notion of sensory preconditioning; e.g., (Brogden 1947; Hall 1996). The small network consisting of these three base level states together with the first-order self-model states representing the characteristics for connection weights and excitability thresholds form a mental model of our subject for the colleague considered here; the connection defines how strongly (in the mental model in the mind of our subject) the face relates to the name of the person. During the scenario these characteristics are learnt so that over time a better mental model and decision result. The network characteristics define the mental model which in terms of Tse (2013) represents the criteria for criterial causation of recognizing a face. Given these criteria, in future situations an encounter with *s* (also at unexpected locations, such as in a shop or in another town) leads to satisfying the criteria and as a consequence to correct recognition.

Note that for this Example Scenario the relevant past pattern *a* can be described by

$$\text{In the past a number of times the face is seen } (\text{srs}_s \text{ occurs}) \text{ co-occurring with the correct belief about which person it was (belief } \text{bs}_i) \tag{16.27}$$

and the relevant future pattern *b* can be described by

$$\text{In the future if the face is seen } (\text{srs}_s, \text{ occurs}), \text{ then the correct recognition takes place (choice } \text{srs}_{a_i}) \tag{16.28}$$

Moreover, the intermediate state property *p* is defined by

$$\text{The adaptive network characteristics have specific appropriate values} \tag{16.29}$$

What such appropriate values are, for example, can be seen in the example simulation in Sect. 16.7; see (16.31).

16.6.2 Connectivity and Aggregation for the Adaptive Network Model

In this section an adaptive network model is introduced that addresses the above example scenario. First, in Fig. 16.6 the connectivity of the base level is depicted. The darker shaded area indicates the mental model considered. After observing the colleague (via ss_s), sensory representation state srs_s of the mental model gets activated, which, depending on the weights of the connections from srs_s to srs_{a_1} and srs_{a_2} and the excitability thresholds of srs_{a_1} and srs_{a_2} ideally activates one of srs_{a_1} and srs_{a_2} indicating the correct colleague, without any of the beliefs bs_1 and bs_2 being activated. However, in the beginning, it is not that ideal: in the first phase, the activation of the correct belief is needed to be able to make a choice between srs_{a_1} and srs_{a_2}. Due to learning, later on this dependence on the beliefs is not needed anymore as the connection from srs_s to the relevant option is strengthened by this learning and the excitability threshold gets lower for that option.

The following two types of self-model states are used to define the adaptation by learning in the considered mental model:

- **Connectivity self-model states for connection weights**
  The states $\mathbf{W}_{X,Y}$ play the role of connection weight for the adaptive connection from X to Y (Hebb 1949)

- **Aggregation self-model states for excitability thresholds**
  The states $\mathbf{T}_Y$ play the combination function parameter role for state Y's adaptive excitability threshold $\boldsymbol{\tau}$ (Chandra and Barkai 2018).

All in all, this creates a core mental model which is a (sub)network based on the following states:

- state srs_s for the image of the face
- states srs_{a_1} and srs_{a_2} for the options of colleagues and their excitability thresholds
- the two connections (dashed arrows) from srs_s to srs_{a_1} and srs_{a_2} for the options of colleagues with their weights
- two connectivity self-model states $\mathbf{W}_{srs_s,srs_{a_1}}$ and $\mathbf{W}_{srs_s,srs_{a_2}}$ for the weights of these two connections
- two aggregation self-model states $\mathbf{T}_{srs_{a_1}}$ and $\mathbf{T}_{srs_{a_2}}$ for the excitability thresholds of states srs_{a_1} and srs_{a_2}

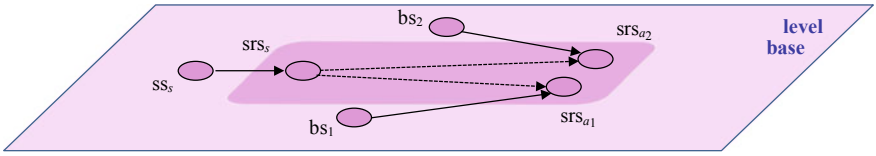


Fig. 16.6 The base level of the example model for recognition

This core mental model can also be extended by adding the belief states to it as well. In a graphical representation of the mental model's *connectivity* in a 3D format, the self-model states are placed in a second (blue) plane, above the (pink) plane for the base network. See Fig. 16.7 and see Table 16.4 for explanations of all states. The following connection types are considered: upward and downward connections, and horizontal leveled connections. Downward connections have a particular effect, as they are effectuating one of the types of adaptive characteristics indicated by their role **W**, **C**, **P** or **H**; see also Table 16.1 in Sect. 16.3.

For *aggregation*, in the example mental model, for the three base level states the logistic function $\mathbf{alogistic}_{\boldsymbol{\sigma},\boldsymbol{\tau}}(\ldots)$ is used, and also for the aggregation (excitability threshold) self-model states $\mathbf{T}_{\mathrm{srs}_{a_1}}$ and $\mathbf{T}_{\mathrm{srs}_{a_2}}$; see (16.17) above. To model Hebbian

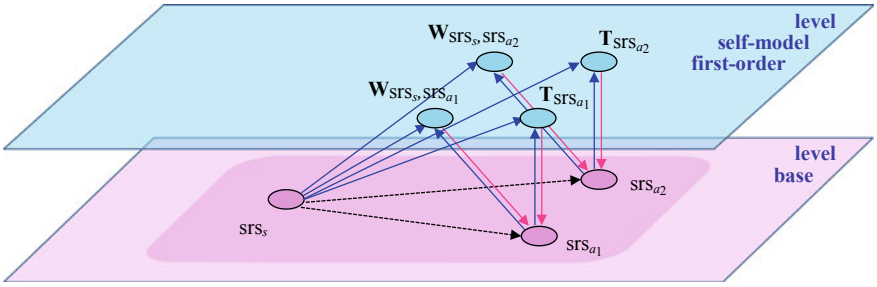


Fig. 16.7. 3D representation of the connectivity of the core mental model model for recognition, to illustrate criterial causation, including: (1) *base level* for the face recognition (depicted by the lower, pink plane), (2) *self-model level* for the citeria (depicted by the upper, blue plane) based on the weights **ω** (for which the two **W**-states are self-model states) of the base connections from srs_s to srs_{a_1} and srs_{a_2} and the excitability thresholds **τ** (for which the **T**-states are self-model states) of these two base states srs_{a_1} and srs_{a_2}

Table 16.4 The states used in the example network model

State nr	name	Explanation
X_1	ss_s	Sensor state for stimulus *s* (seeing a face)
X_2	srs_s	Sensory representation state for stimulus *s*
X_3	bs_1	Belief state 1 (belief that it is Person 1)
X_4	bs_2	Belief state 2 (belief that it is Person 2)
X_5	srs_{a_1}	Sensory representation state for recognition as Person 1
X_6	srs_{a_2}	Sensory representation state for recognition as Person 2
X_7	$\mathbf{W}_{\mathrm{srs}_s,\mathrm{srs}_{a_1}}$	Self-model state for the weight of the connection from srs_s to srs_{a_1}
X_8	$\mathbf{W}_{\mathrm{srs}_s,\mathrm{srs}_{a_2}}$	Self-model state for the weight of the connection from srs_s to srs_{a_2}
X_9	$\mathbf{T}_{\mathrm{srs}_{a_1}}$	Self-model state for the excitability threshold of srs_{a_1}
X_{10}	$\mathbf{T}_{\mathrm{srs}_{a_2}}$	Self-model state for the excitability threshold of srs_{a_2}

learning (Hebb 1949), combination function $\mathbf{hebb}_{\mu}(V_1, V_2, W)$ is applied for the two connectivity (connection weight) self-model states $\mathbf{W}_{X,Y}$ within the mental model

$$\mathbf{hebb}_{\mu}(V_1, V_2, W) = V_1 V_2 (1 - W) + \mu W \tag{16.30}$$

where V_1, V_2 are variables for the single impacts the self-model state $\mathbf{W}_{X,Y}$ (in Fig. 16.5 in the upper plane) gets from the base states X and Y in the addressed adaptive connection (in Fig. 16.7 in the lower plane) and W for the connection weight representation $\mathbf{W}_{X,Y}$; moreover, $\boldsymbol{\mu}$ is a persistence parameter.

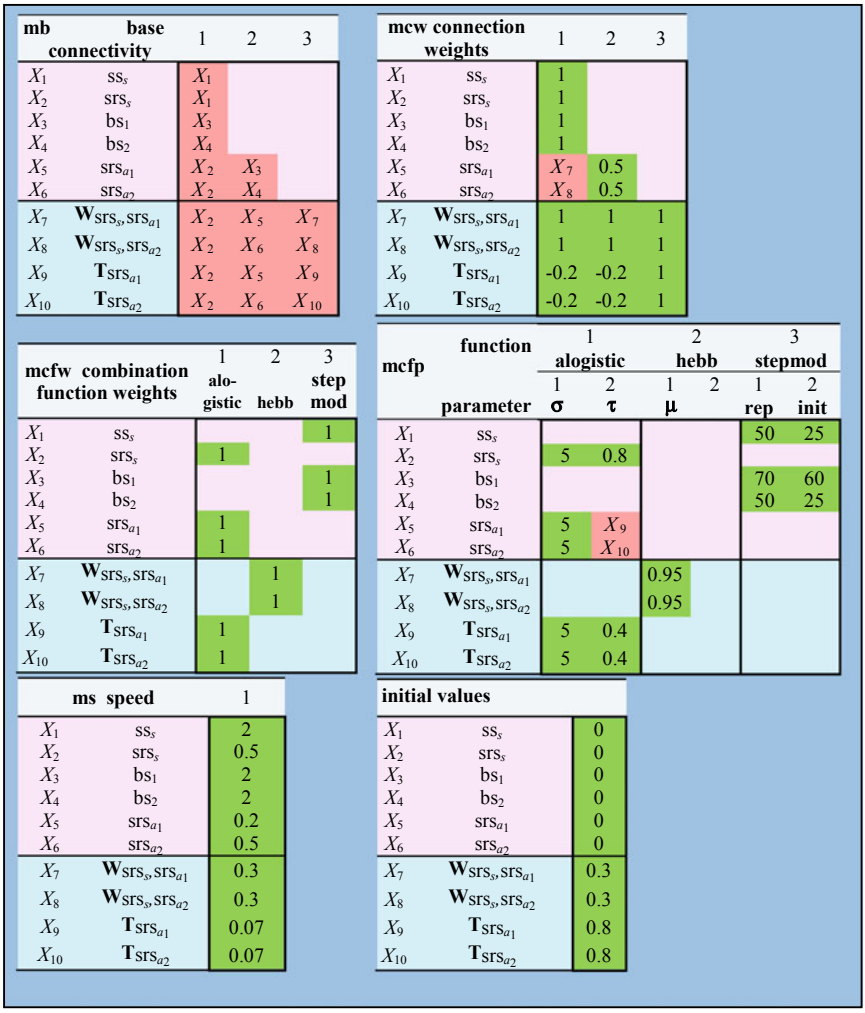
16.6.3 Specification of the Adaptive Network Model by Role Matrices

The network model was specified by *role matrices*. They are **mb** (for the base connection role), **mcw** (for the connection weight role), **ms** (for the speed factor role), **mcfw** (for the combination function weight role), and **mcfp** (for the combination function parameter role); e.g., (Treur 2020a, b). In all of these role matrices (see Fig. 16.8) the rows are for the different states; at each row for the state X_j indicated in the first column, it is specified in the other columns which other states (via the incoming arrows shown in Fig. 16.7) or values have impact on X_j for that role.

These impacts on X_j are distinguished by their role, i.e.: *base* or *non-base*, where for the non-base case a distinction is made for respectively: *connection weight role*, *speed factor role*, *combination function weight role* and *combination function parameter role*. For a designed model a list of combination functions used is specified by **mcf** = [….], for the current example it is **mcf** = [2 3 35]; here the numbers 2, 3, 35 refer to the numbers in the combination function library, where $\mathbf{alogistic}_{\sigma,\tau}(\ldots)$ has number 2 and $\mathbf{hebb}_{\mu}(\ldots)$ number 3; number 35 is the stepmod function used to create independent events. Box 16.2 shows all role matrices for the adaptive network model addressing criterial causation.

Role matrix **mb** specifying *base connectivity* indicates at each row for the indicated state X_j from which states it gets incoming connections from the same or a lower level. For example, the 5th row indicates for state X_5 (= srs_{a_1}) two incoming base connections, one from state X_2 (= srs_s), and one from state X_3 (= bs_1). For another example, row 7 indicates that state X_7 (= $\mathbf{W}_{\mathrm{srs}_s,\mathrm{srs}_{a_1}}$) has incoming base connections from X_2 (= srs_s), X_5 (= srs_{a_1}) and from X_7 itself in that order; this ordering is crucial since the Hebbian combination function $\mathbf{hebb}_{\mu}(\ldots)$ used for this state X_7 (= $\mathbf{W}_{\mathrm{srs}_s,\mathrm{srs}_{a_1}}$) is not symmetric in its three arguments, as can be seen in (16.30).

The four role matrices specifying *non-base connectivity* describe in each row impacts on the indicated state X_j by self-model states from a higher level (see the downward arrows in Fig. 16.7); they are as follows: role matrices **mcw** for the connection weight role and **ms** for the speed factor role, and role matrices **mcfw**



mb	base connectivity	1	2	3
X_1	ss_s	X_1		
X_2	srs_s	X_1		
X_3	bs_1	X_3		
X_4	bs_2	X_4		
X_5	srs_{a_1}	X_2	X_3	
X_6	srs_{a_2}	X_2	X_4	
X_7	$\mathbf{W}srs_s,srs_{a_1}$	X_2	X_5	X_7
X_8	$\mathbf{W}srs_s,srs_{a_2}$	X_2	X_6	X_8
X_9	$\mathbf{T}srs_{a_1}$	X_2	X_5	X_9
X_{10}	$\mathbf{T}srs_{a_2}$	X_2	X_6	X_{10}

mcw	connection weights	1	2	3
X_1	ss_s	1		
X_2	srs_s	1		
X_3	bs_1	1		
X_4	bs_2	1		
X_5	srs_{a_1}	X_7	0.5	
X_6	srs_{a_2}	X_8	0.5	
X_7	$\mathbf{W}srs_s,srs_{a_1}$	1	1	1
X_8	$\mathbf{W}srs_s,srs_{a_2}$	1	1	1
X_9	$\mathbf{T}srs_{a_1}$	-0.2	-0.2	1
X_{10}	$\mathbf{T}srs_{a_2}$	-0.2	-0.2	1

mcfw	combination function weights	1 alo-gistic	2 hebb	3 step mod
X_1	ss_s			1
X_2	srs_s	1		
X_3	bs_1			1
X_4	bs_2			1
X_5	srs_{a_1}	1		
X_6	srs_{a_2}	1		
X_7	$\mathbf{W}srs_s,srs_{a_1}$		1	
X_8	$\mathbf{W}srs_s,srs_{a_2}$		1	
X_9	$\mathbf{T}srs_{a_1}$	1		
X_{10}	$\mathbf{T}srs_{a_2}$	1		

mcfp	function	1 alogistic		2 hebb		3 stepmod	
	parameter	1 σ	2 τ	1 μ	2	1 rep	2 init
X_1	ss_s					50	25
X_2	srs_s	5	0.8				
X_3	bs_1					70	60
X_4	bs_2					50	25
X_5	srs_{a_1}	5	X_9				
X_6	srs_{a_2}	5	X_{10}				
X_7	$\mathbf{W}srs_s,srs_{a_1}$			0.95			
X_8	$\mathbf{W}srs_s,srs_{a_2}$			0.95			
X_9	$\mathbf{T}srs_{a_1}$	5	0.4				
X_{10}	$\mathbf{T}srs_{a_2}$	5	0.4				

ms	speed	1
X_1	ss_s	2
X_2	srs_s	0.5
X_3	bs_1	2
X_4	bs_2	2
X_5	srs_{a_1}	0.2
X_6	srs_{a_2}	0.5
X_7	$\mathbf{W}srs_s,srs_{a_1}$	0.3
X_8	$\mathbf{W}srs_s,srs_{a_2}$	0.3
X_9	$\mathbf{T}srs_{a_1}$	0.07
X_{10}	$\mathbf{T}srs_{a_2}$	0.07

initial values		
X_1	ss_s	0
X_2	srs_s	0
X_3	bs_1	0
X_4	bs_2	0
X_5	srs_{a_1}	0
X_6	srs_{a_2}	0
X_7	$\mathbf{W}srs_s,srs_{a_1}$	0.3
X_8	$\mathbf{W}srs_s,srs_{a_2}$	0.3
X_9	$\mathbf{T}srs_{a_1}$	0.8
X_{10}	$\mathbf{T}srs_{a_2}$	0.8

Fig. 16.8 Role matrices specification for the example network addresing criterial causation

for the combination function weight role and **mcfp** for the combination function parameter role (see Fig. 16.8). Within each of these non-base role matrices cell entries in red cells show the name of a state (at a higher level) that as self-model state represents in an adaptive characteristic; in contrast, entries in green cells indicate static values for nonadaptive characteristics. Therefore, as seen in Fig. 16.8 the red cells in **mcw** and **mcfp** indicate the (self-model) states X_7 to X_{10}. For example, in role matrix **mcw** the indication X_7 in the red cell at row 5 and column 1 specifies that the value of state X_7 represents the connection weight from srs_s to srs_{a_1} (as indicated in **mb**). Unlike this, the 1 in green cell at row 7, column 1 of **mcw** shows the nonadaptive value of weight of the connection from X_2 (= srs_s) to X_7 (= $\mathbf{W}_{srs_s,srs_{a_1}}$). In role matrix

mcfp specifying the combination function parameter role, in the red cell at row 5 and column 2 it is specified that the actual value for the excitability threshold of srs_{a_1} is represented by the value of self-model state X_9 ($= \mathbf{T}_{\text{srs}_{a_1}}$). More explanation of this specification format and how it is used to automatically generate simulations can be found in (Treur 2020a, b).

16.7 Simulation of the Development and Use of the Mental Model

This section describes one of the example simulation scenarios for the mental model described in Sect. 16.6 that were addressed using the modeling environment developed (Treur 2020a, b). In particular, Fig. 16.9 shows a simulation for the Example Scenario described in Sect. 16.6. The settings from Fig. 16.8 were used. In accordance with temporal factorisation, past pattern a and future pattern b, and the mediating state property p, will be discussed subsequently.

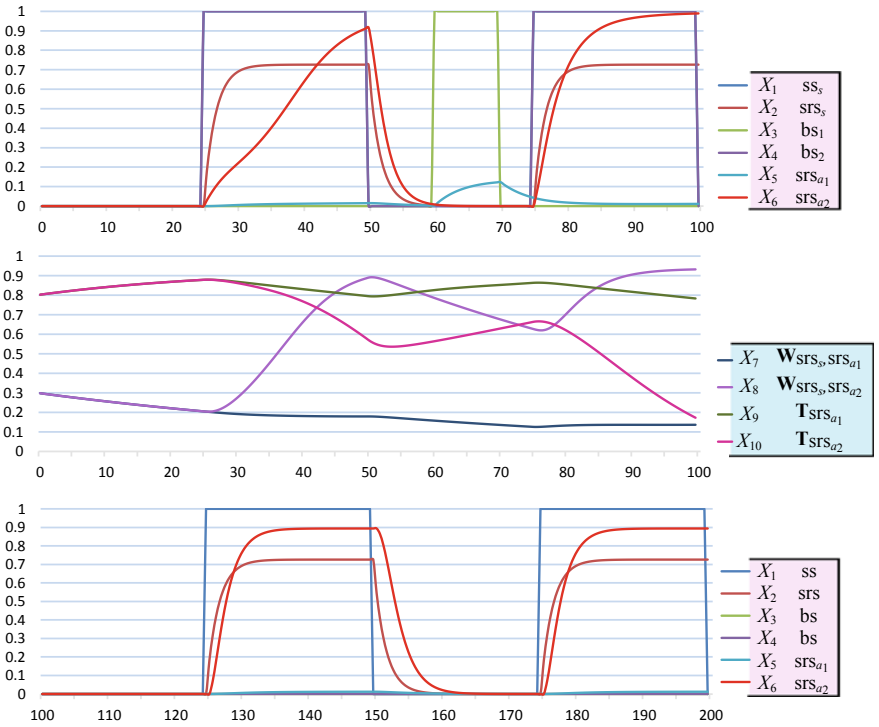


Fig. 16.9 The past trace (shown in upper graph + middle graph), and the future trace (shown in the lower graph). Here the criterion as a mediating state (set by the past pattern) is driving the future pattern

16.7.1 Past Pattern a (Time Point 0 to 100)

In the past trace according to pattern a, the stimulus s (indicating the observed face) is active from time point 25 to time point 50 and from time 75 to time 100. In these time periods, the belief state bs_2 co-occurs. In Fig. 16.9, the upper and middle graph show this past trace and how recognition of the stimulus s as Person 2 improves over time: during the first encounter, the mental model state for sensory representation srs_{a_2} (the red line) only increases slowly, but during the second encounter this happens much faster; apparently at that time a better criterion was set for proper recognition already.

The middle graph shows the emerging network characteristics representing this criterion. It can be seen that (under influence of belief state bs_2):

- the adaptive connection weight from srs_s to srs_{a_2} represented by self-model state $\text{W}_{\text{srs}_s,\text{srs}_{a_2}}$ (purple line) of the mental model gets stronger, as a form of sensory preconditioning (Brogden 1947; Hall 1996) by a Hebbian learning mechanism (Hebb 1949)
- the excitability threshold represented by $\text{T}_{\text{srs}_{a_2}}$ of srs_{a_2} (pink line) gets lower, which relates to the type of intrinsic excitability adaptation described more in general in (Chandra and Barkai 2018).

Between time points 60 and 70 as an intended irregularity of the history, for a very short period, belief state bs_1 shows, but this has no serious consequences for the adaptation process.

16.7.2 Criterion Formed at Time Point 100

Mediating state property p in general describes the general criterion (16.15) for criterial causation. In the addressed scenario, the considered mediating state property p describes the considered mental model in terms of its network characteristics: the self-model states for the weight representations $\mathbf{W}_{\text{srs}_s,\text{srs}_{a_1}}$ and $\mathbf{W}_{\text{srs}_s,\text{srs}_{a_2}}$ of the connections from srs_s to srs_{a_1} and srs_{a_2}, and the self-model states for the excitability thresholds $\mathbf{T}_{\text{srs}_{a_1}}$ and $\mathbf{T}_{\text{srs}_{a_2}}$ for srs_{a_1} and srs_{a_2}; so mediating state property p covers the values of self-model states X_7 to X_{10} of the mental model. Note that all belong to the self-model level.

Time point 100 shows the following assignments of values: $X_7 = 0.136275$, $X_8 = 0.93172$, $X_9 = 0.78281$, $X_{10} = 0.17232$. It is in particular this configuration (at the self-model level) that defines the mediating state property p:

$$p \equiv X_7 = 0.136275 \quad \text{and} \quad X_8 = 0.93172 \quad \text{and} \quad X_9 = 0.78281 \quad \text{and} \quad X_{10} = 0.17232 \tag{16.31}$$

In this way p describes the values for $\boldsymbol{\omega}_{X_i,Y}$ and $\boldsymbol{\tau}$ in criterion (16.26)

$$\boldsymbol{\omega}_{X_1,Y} X_1(t) + \cdots + \boldsymbol{\omega}_{X_k,Y} X_k(t) > \mathrm{c}$$

$$\text{with } \mathrm{c} = \boldsymbol{\tau} - \log\left(\frac{1}{\frac{0.5}{1+\mathrm{e}^{-\sigma\,\tau}} + \frac{1}{1+\mathrm{e}^{\sigma\,\tau}}} - 1\right)/\sigma$$

for the advanced logistic function for criterial causation.

At time 100, representation $\mathbf{W}_{\mathrm{srs}_s,\mathrm{srs}_{a_1}}$ for the weight of the considered connection from stimulus representation srs_s to sensory representation srs_{a_1} within the mental model has a low value (0.136275), whereas representation $\mathbf{T}_{\mathrm{srs}_{a_1}}$ for the excitability threshold $\boldsymbol{\tau}_{\mathrm{srs}_{a_1}}$ for srs_{a_1} within the mental model has a high value (0.78281). This makes that for Person 1 the criterion for activation of srs_{a_1} within the mental model cannot be fulfilled.

For sensory representation state srs_{a_2} of the mental model it is completely different: at time 100 the representation $\mathbf{W}_{\mathrm{srs}_s,\mathrm{srs}_{a_2}}$ for the weight of the concerning connection from stimulus representation srs_s to sensory representation srs_{a_2} within the mental model has a high value (0.93171758), whereas representation $\mathbf{T}_{\mathrm{srs}_{a_2}}$ for the excitability threshold $\boldsymbol{\tau}_{\mathrm{srs}_{a_2}}$ for srs_{a_2} within the mental model has a low value (0.17232). Therefore, the criterion for Person 2 can be fulfilled easily just by the impact only coming from srs_s. Indeed, using (16.26) after substitution by the values at time point 100 of the relevant **W**- and **T**-states (which represent **ω** and **τ**), and substitution of the static value 5 for **σ**, and 0.5 for the static connection weight **ω** from the belief state, the criterion for activation translates into

$$0.5\ \mathrm{bs}_2(t) + 0.93172\ \mathrm{srs}_s(t) > 0.0857$$

No positive value of $\mathrm{bs}_2(t)$ is needed here; this is fulfilled if

$$\mathrm{srs}_s(t) > 0.0857/0.93172 = 0.092$$

This implies that already a low-level sensory face representation signal srs_s (as low as 0.1) is enough for the face recognition based on the acquired mental model.

16.7.3 Future Pattern b (Time Point 100–200)

The lower graph shown in Fig. 16.9 depicts the future trace according to pattern *b*; it can be seen how based on the learnt mental model recognition takes place without any activation of the belief state. This mental model defines the criterion in terms of the mediating state property at time 100. The criterion is described by the mental model defined by the value assignments for the self-model states for connection weights and for the excitability thresholds shown in (16.31) in Sect. 7.2 above. In the future trace at times 125 and 175 the face is encountered again (ss_s is activated),

whereby the corresponding belief state bs_2 is kept 0 (so no additional circumstantial info available), and the criterion is satisfied so that adequate recognition srs_{a_2} as Person 2 occurs.

16.8 Defining Informational Content by Temporal Relational Specification

This section addresses the question how exactly the mediating state for temporal factorisation and, equivalently for the case of the brain, the state of the criteria for criterial causation, can be interpreted as having informational meaning. In other words: how can the informational content of these states be defined? Here something can be learned from Philosophy of Mind.

16.8.1 Relational Specification of Mental Content

From Philosophy of Mind, well-known philosopher Jaegwon Kim (1996), has put forward the notion of relational specification, in the context of defining mental content of (mental or physical) state properties:

> 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. (Kim 1996), pp. 200–201; italics in the original

For the temporal perspective, in particular, a temporal relational specification is the specification of temporal relationships ('both current and historical') of the considered state to certain patterns in past and/or future. Kim puts forward that relational specifications are crucial to be able to express laws and explanations.

> Consider physical magnitudes such as mass and length, which are standardly considered to be paradigm examples of intrinsic properties of material objects. But how do we *specify*, *represent*, or *measure* the mass or length of an object? The answer: relationally. To say that this rod has a mass of 5 kilograms is to say that it bears a certain relationship to the International Prototype Kilogram (it would balance, on an equal-arm balance, five objects each of which balances the Standard Kilogram). Likewise, to say that the rod has a length of 2 meters is to say that it is twice the length of the Standard Meter (or twice the distance travelled by light in a vacuum in a certain specified fraction of a second). These properties are intrinsic, but their specifications or representations are extrinsic and relational, involving relationships to other things and properties in the world. It may well be that the availability of such extrinsic representations are essential to the utility of these properties in the formulation of scientific laws and explanations. (Kim 1996, p. 201); italics in the original

In Kim's approach, a considered (mental) state is distinct from the relationships it has to external other items (Kim 1996, pp. 200–202). Kim explains as follows that a state property itself is intrinsic, whereas its relational specification indicates how it is linked to items in the world:

> 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. 201–202)

16.8.2 Applying Temporal Relational Specification to Informational Content

The above summarized notion of relational specification for mental states, provides a way to specify the informational content of mediating state property p postulated by temporal factorisation. For the future direction, the occurrence of p at some time t makes b occur in the future after t. A temporal relational specification may express what the influence of p on the future is, namely that pattern b will occur, also indicated by $p \Rightarrow b$. In a similar manner a temporal relational specification with respect to the past can be described by $a \Rightarrow p$. Then in fact the basic part of the consequent of the principle of temporal factorisation

$$a \Rightarrow p \;\; \& \;\; p \Rightarrow b \tag{16.32}$$

can be identified as a temporal relational specification of the informational content of state p.

For the scenario described by the example network model with connectivity depicted in Fig. 16.1, the following two temporal statements are used to describe patterns a and b for this example which more informally described by (16.11) and (16.12) for the locked door example, and by (16.20) and (16.21) for the delayed response animal behaviour example:

- Temporal statement expressing pattern a

$$\begin{aligned}&\text{For some point } t_1 < t \text{ it holds } X_1(t_1) > 0.5 \text{ and for al time points } t_2\\ &\text{with } t_1 < t_2 < t \text{ it does not hold } X_1(t_1) \leq 0.5\end{aligned} \tag{16.33}$$

- Temporal statement expressing pattern b

$$\begin{aligned}&\text{If for some point } t_3 > t \text{ it holds } X_3(t_3) > 0.5\\&\text{then there is some time point } t_4 > t_3 \text{ such that it holds } X_4(t_4) > 0.5 \quad (16.34)\end{aligned}$$

Note that here, to avoid more complex temporal expressions, only an approximation is discussed. To get a more precise treatment, certain minimal durations have to be specified, since a state that has a value > 0.5 for just one single time point has not much effect. This means that a specific trace over time (which is a sequence of state values for all states for all time points, like in a simulation outcome) satisfies the pattern if and only if the corresponding temporal statement is satisfied by (or valid for) this trace. Note that these two temporal expressions (16.33) and (16.34) are indeed valid for the simulation trace shown in Fig. 16.4 when t is set between 42 and 46. Given the above pattern descriptions, the informational content of intermediate state property X_2 can be illustrated in a simple version as follows.

- Informational content of X_2 for the past:

$$\begin{aligned}&\text{If for some point } t_1 < t \text{ it holds } X_1(t_1) > 0.5 \text{ and for all time points } t_2\\&\text{with } t_1 < t_2 < t \text{ it does not hold } X_1(t_1) \leq 0.5 \quad \text{then } X_2(t) > 0.5 \quad (16.35)\end{aligned}$$

- Informational content of X_2 for the future:

$$\begin{aligned}&\text{If } X_2(t) > 0.5 \text{ then}\\&\text{if for some point } t_3 > t \text{ it holds } X_3(t_3) > 0.5\\&\text{then there is some time point } t_4 > t_3 \text{ such that it holds } X_4(t_4) > 0.5 \quad (16.36)\end{aligned}$$

Note that these temporal statements (16.35) and (16.36) are satisfied (are valid) for the simulation trace shown in Fig. 16.4 when t is set between 42 and 46. This notion of informational content by temporal relational specification can loosely be formulated as: the informational content of p is the occurrence of pattern a in the past and of pattern b in the future. Note that here it might be more satisfactory if it is more explicitly formulated that the occurrence of past pattern a is also implied by p, and the occurrence of p is implied by future pattern b, which actually is a kind of converse of (16.32). Then (16.32) can be sharpened to

$$a \Leftrightarrow p \;\;\&\;\; p \Leftrightarrow b \qquad (16.37)$$

This can be formulated as: the informational content of p is on the one hand the occurrence of pattern a in the past and on the other hand of pattern b in the future. Then for the above example it becomes:

- Informational content of X_2 for the past:

$$\begin{aligned}&\text{For some point } t_1 < t \text{ it holds } X_1(t_1) > 0.5 \text{ and for all time points } t_2\\ &\text{with } t_1 < t_2 < t \text{ it does not hold } X_1(t_1) \leq 0.5\\ &\text{if and only if } X_2(t) > 0.5\end{aligned} \tag{16.38}$$

- Informational content of X_2 for the future:

$$\begin{aligned}&X_2(t) > 0.5 \text{ if and only if}\\ &\quad \text{if for some point } t_3 > t \text{ it holds } X_3(t_3) > 0.5\\ &\quad \text{then there is some time point } t_4 > t_3 \text{ such that it holds } X_4(t_4) > 0.5\end{aligned} \tag{16.39}$$

Note that again these temporal statements (16.38) and (16.39) are satisfied (are valid) for the simulation trace shown in Fig. 16.4 when t is set between 42 and 46. The latter if-and-only-if approach expressed in (16.38) and (16.39) might fit better to an intuition of representational content.

For the example mental model for recognition of the colleague described in Sects. 16.6 and 16.7, similarly the relational specification of the past and future informational content of this mental model roughly spoken are as follows.

- Informational content of the mental model for the past:

A number of times the face was observed while simultaneously
the belief was activated that it was Person 2

- Informational content of the mental model for the future:

If at some later point in time the face is observed then it is Person 2

Note that the informational content of the mental model for the future is indeed precisely the informational content assigned by Tse (2013) as criteria for future firing of a neuron. So, these two notions on relational specification of informational content for the future according to Kim (1996) and Tse (2013)'s informational interpretation of the current state as criteria for future firing are just equivalent as ways of assigning informational content to the mental model.

In fact, the difference between (16.38) and (16.39) and (16.35) and (16.36) lies in an additional *temporal completion assumption*. Such an assumption indicates that the description of pattern a covers *all* possibilities in the past to get p, and similarly, state description p covers *all* possibilities for b to occur in the future. Adopting this additional temporal completion assumption, the stronger (16.38) and (16.39) can be taken as describing the informational content of p. Note that under this temporal completion assumption, also the temporal factorisation principle can get a different

form, where in addition to (16.37), also by temporal completion $a \Rightarrow b$ is replaced by $a \Leftrightarrow b$:

$$[a \Leftrightarrow b] \Rightarrow \exists p[a \Leftrightarrow p \quad \& \quad p \Leftrightarrow b] \tag{16.40}$$

There are pros and cons for this temporal completion assumption. An interesting pro is that more elegant equivalences are obtained for the informational content of p. A con may be that it might lead to disjunctive expressions for the past and future pattern specifications and for p itself to cover all alternative possibilities; this may complicate things and may not always be applicable well.

16.9 Formalisation of Temporal Factorisation and Criterial Causation in Temporal Trace Predicate Logic

In this section it is shown how the notions discussed in the chapter can be formalised in terms of the formal language of a form of Predicate Logic called Temporal Trace Predicate Logic TTPL. This is a use of Predicate Logic for which numbers can be used and for which constants, terms and variables can be used for time points t, state properties p, and traces tr. For somewhat similar or comparable uses of Predicate Logic see (Bosse et al. 2009; Bosse and Treur 2011; Sharpanskykh and Treur 2010; Treur 2009) or (Treur 2016), Chap. 13. Here it is discussed for Temporal Factorisation, but as Criterial Causation is considered to be a special case of Temporal Factorisation it applies to Criterial Causation as well.

16.9.1 Formalisation of Temporal Factorisation in Temporal Trace Predicate Logic

First the notions of past pattern and future pattern have to be formalised:

Past and future trace properties and current properties

A *past trace property* is a property *P(t, tr)* for time point t and trace tr for which every quantor for a variable t' over time is relativised by $t' < t$

A *future trace property* is a property *F(t, tr)* for time point *t* and trace *tr* for which every variable *t'* over time is relativised by $t' > t$

A *current trace property* C(*p*, *t*, *tr*) about the present expresses that in trace *tr* at *t* state property *p* holds

These past trace properties and future trace properties are used to describe past patterns and future patterns for the indicated trace *tr*. Then the following formalisation for temporal factorisation is obtained.

In Box 16.2 the formalisation of Temporal Factorisation in TTPL is shown.Here the predicate Stateproperty indicates the state properties used.

Box 16.2 Temporal Factorisation formalised in TTPL.

For every past trace property *P(t, tr)* and future trace property *F(t, tr)* it holds

$$\forall t, tr[P(t, tr) \Rightarrow F(t, tr)] \Rightarrow$$
$$\exists p[\text{Stateproperty}(p) \& \forall t, tr[[P(t, tr) \Rightarrow \text{C}(p, t, tr)] \& [\text{C}(p, t, tr) \Rightarrow F(t, tr)]]]$$

An illustration for the lock and mouse example of Sect. 16.4 is shown in Box 16.3; see also (16.21) and (16.22) in Sect. 16.8. Note that it x is a state, then *X(t, tr)* denotes the value of *X* at time *t* in trace *tr*.

Box 16.3 Temporal Factorisation for the example of Sect. 16.4 formalised in TTPL

$$P(t, tr) \equiv \exists t_1 < t\ X_1(t_1, tr) > 0.5 \& \forall t_2 > t_1[t_2 < t \Rightarrow \text{not}\ X_1(t, tr) \leq 0.5]$$
$$F(t, tr) \equiv \forall\ t_3 > t\ X_3(t_3, tr) > 0.5 \Rightarrow \exists t_4 > t_3 X_4(t_4, tr) > 0.5$$
$$\text{C}(p, t, tr) \equiv X_2(t, tr) > 0.5$$

Temporal factorisation for this case for state property *p* formalised as C(*p, t, tr*) :

$$[\forall t, tr[[\exists t_1 < t\ X_1(t_1, tr) > 0.5 \& \forall t_2 > t_1[t_2 < t \Rightarrow \text{not}\ X_1(t_1, tr) \leq 0.5]] \Rightarrow$$
$$[\forall t_3 > t\ X_3(t_3, tr) > 0.5 \Rightarrow \exists t_4 > t_3 X_4(t_4, tr) > 0.5]]] \Rightarrow$$

$$[\forall t, tr[[\exists t_1 < t\ X_1(t_1, tr) > 0.5 \& \forall t_2 > t_1[t_2 < t \Rightarrow \text{not} X_1(t_1, tr) \leq 0.5] \Rightarrow X_2(t, tr) > 0.5] \&$$
$$[X_2(t, tr) > 0.5 \Rightarrow \forall t_3 > t\ X_3(t_3, tr) > 0.5 \Rightarrow \exists t_4 > t_3\ X_4(t_4, tr) > 0.5]]]$$

Informational Content by temporal relational specification

The statements (16.35) and (16.36) from Sect. 16.7 defining Informational Content can be formalised as shown in Box 16.4.

Box 16.4 Informational content formalised in TTPL by implications and illustrated for the example of Sect. 16.4.

$$\forall t, tr[P(t, tr) \Rightarrow \mathrm{C}(p, t, tr)]$$
$$\forall t, tr[\mathrm{C}(p, t, tr) \Rightarrow F(t, tr)]$$

For the example of Sect. 16.4:

$$\forall t, tr[[\exists t_1 < t\ X_1(t_1, tr) > 0.5 \& \forall t_2 > t_1 \quad [t_2 < t \Rightarrow \text{not } X_1(t_1, tr) \leq 0.5]$$
$$\Rightarrow X_2(t, tr) > 0.5]$$
$$\forall t, tr[X_2(t, tr) > 0.5 \Rightarrow \forall t_3 > t\ X_3(t_3, tr) > 0.5 \Rightarrow \exists t_4 > t_3\ X_4(t_4, tr) > 0.5]$$

The statements (16.38) and (16.39) from Sect. 16.7 defining Informational Content by bi-implications can be formalised as shown in Box 16.5.

Box 16.5 Informational content formalised in TTPL by bi-implications as illustrated for the example of Sect. 16.4.

$$\forall t, tr[P(t, tr) \Leftrightarrow \mathrm{C}(p, t, tr)]$$
$$\forall t, tr[\mathrm{C}(p, t, tr) \Leftrightarrow F(t, tr)]$$

For the example of Sect. 16.4:

$$\forall t, tr[[\exists t_1 < t\ X_1(t_1, tr) > 0.5 \& \forall t_2 > t_1 \quad [t_2 < t \Rightarrow \text{not } X_1(t_1, tr) \leq 0.5] \Leftrightarrow X_2(t, tr) > 0.5]$$
$$\forall t, tr[X_2(t, tr) > 0.5 \Leftrightarrow \forall t_3 > t[X_3(t_3, tr) > 0.5 \Rightarrow \exists t_4 > t_3\ X_4(t_4, tr) > 0.5]]$$

As it has been shown that assigning informational content by Kim (1996)'s future-directed relational specification is equivalent to Tse (2013)'s way of assigning informational content as criterial causation, the last line in Box 16.5 is also a formalisation of this notion of criterial causation in TTPL.

16.9.2 Formalisation of Temporal Factorisation in Reified Temporal Trace Predicate Logic

It is in principle also possible to use a reified predicate logic RTTPL in which the temporal properties expressed as statements of TTPL are reified by names for which constants, terms and variables can be used. For a similar approach, see MetaTTL in (Bosse and Treur 2011). Then temporal factorisation can be formalised as shown in Box 16.6.

Box 16.6 Temporal Factorisation formalised in RTTPL.

$$\begin{aligned}&\forall P, F\ [\ [\ \text{Pasttraceproperty}(P, t)\ \&\ \text{Futuretraceproperty}(F, t)\ \&\\ &\qquad \forall t, tr\ [\text{Holdsfor}(P, t, tr) \Rightarrow \text{Holdsfor}(F, t, tr)]] \Rightarrow\\ &\qquad\quad \exists p\ [\ \text{Stateproperty}(p)\ \&\\ &\qquad\qquad \forall t, tr\ [\ [\text{Holdsfor}(P, t, tr) \Rightarrow \text{Holdsat}(p, t, tr)]\ \&\\ &\qquad\qquad\qquad [\text{Holdsat}(p, t, tr) \Rightarrow \text{Holdsfor}(F, t, tr)]]]].\end{aligned}$$

Here predicates Past trace property (P, t) and Future trace property (F, t) are used to indicate these properties, Holdsfor(X, t, tr) expresses that temporal property X is satisfied for trace tr with respect to time point t, and Holdsat(Y, t, tr) that state property Y is valid for time t in trace tr. Although in principle this works, it is technically much more complicated compared to TTPL, as the coding of temporal properties requires some technical notations.

16.10 Discussion

To clarify how a mental model has informational content, the notion of criterial causation as introduced in (Tse 2013) was considered. This notion describes how in the brain by plasticity specific configurations can emerge that represent informational criteria for future processing. The current chapter addressed how this notion, which can be considered a special case of the notion of temporal factorisation for world dynamics introduced in (Treur 2007a), applies to learning and use of mental models. It is mainly based on material from (Treur 2021).

The core of the idea is that according to some adaptive process, (past) brain patterns or world patterns occur, which as patterns lead to emerging brain configurations or world configurations in the present; these configurations in turn drive or affect the (future) brain pattern or world pattern. Such a configuration, in one (present) state encodes information about the past which is relevant for the future. Future processes are driven by this information (for a person, or for the world). This idea was applied to mental models by considering a mental model as representing this emerging brain configuration. It was shown how an architecture for dynamics and adaptation of

mental models generating these processes, can be related to these notions of criterial causation and temporal factorisation.

More specifically, it was discussed how a mental model that is learnt can be considered to have informational content like criteria for criterial causation, and how from the perspective of (Kim 1996) its informational content can be described by relational specification both for the past and for the future. Here it has been shown that the latter (future) type of relational specification is equivalent to Tse (2013)'s interpretation of the mental model as criteria for future activation of a considered state.

For future work, it can explored in how far the notion of Extended Mind (e.g., Bosse et al. 2005; Bosse et al. 2005; Clark and Chalmers 1998) can be covered in a similar manner, where both the brain and the world play their role. This goes beyond the notion of criterial causation (which was only described for brain states), but still fits in the more general temporal factorisation principle. Applied in particular for mental models, this means that the mental model or part of it is represented in the world and not in the brain; for example, a complex flow chart representing a mental model of some process as displayed on a computer screen. Moreover, it can be explored how also metaplasticity (Abraham and Bear 1996; Treur 2020a) can be addressed in adaptive processes concerning criterial causation, that will become second-order adaptive then.

In the current chapter, links to the notions free will or mental causation (as discussed by Tse) from Philosophy of Mind has been left aside; this was on purpose. The notion of criterial causation of Tse (2013) has much value by itself, independent of such philosophical links, as is also emphasized by Levy (2013). This value has been the focus here.

References

Abraham, W.C., Bear, M.F.: Metaplasticity: the plasticity of synaptic plasticity. Trends Neurosci. **19**(4), 126–130 (1996)

Ashby, W.R.: Design for a Brain. Chapman & Hall London Revised edition 1960 (1952)

Bosse, T., Jonker, C.M., van der Meij, L., Sharpanskykh, O., Treur, J.: specification and verification of dynamics in agent models. Int. J. Coop. Inf. Syst. **18**, 167–193 (2009)

Bosse, T., Jonker, C.M., Schut, M.C., Treur, J.: Simulation and analysis of shared extended mind. Simul. J. (soc. Model. Simul.) **81**, 719–732 (2005)

Bosse, T., Jonker, C.M., Schut, M.C., Treur, J.: Collective representational content for shared extended mind. Cogn. Syst. Res. **7**, 151–174 (2006)

Bosse, T., Treur, J.: Patterns in world dynamics indicating agency. Trans. Comput. Collective Intell. **3**, 128–151 (2011)

Brogden, W.J.: Sensory preconditioning of human subjects. J. Exp. Psychol. **37**, 527–539 (1947)

Chandra, N., Barkai, E.: A non-synaptic mechanism of complex learning: modulation of intrinsic neuronal excitability. Neurobiol. Learn. Mem. **154**, 30–36 (2018)

Clark, A., Chalmers, D.: The extended mind. Analysis **58**, 7–19 (1998)

Cromwell, H.C., Tremblay, L., Schultz, W.: Neural encoding of choice during a delayed response task in primate striatum and orbitofrontal cortex. Exp. Brain Res. **236**(6), 1679–1688 (2018)

Descartes, R.: The World Ch 6: Description of a New World and on the Qualities of the Matter of Which it is Composed (1634)
Foster, J.M.: Unit activity in the prefrontal cortex during delayed response performance: neuronal correlates of short-term memory. J. Neurophysiol. **36**, 61–78 (1973)
Hall, G.: Learning about associatively activated stimulus representations: implications for acquired equivalence and perceptual learning. Anim. Learn. Behav. **24**, 233–255 (1996)
Hebb, D.O.: The Organization of Behavior: A Neuropsychological Theory. Wiley NY (1949)
Hunter, W.S.: The delayed reaction in animals. Behav. Monogr. **2**, 1–85 (1912)
Kim, J.: Philosophy of Mind Westview Press (1996)
Laplace, P.S.: Philosophical Essays on Probabilities. Springer-Verlag New York 1995 Translated by AI Dale from the 5th French edition of 1825 (1825)
Levy, N.: Review of P.U. Tse—the neural basis of free will: criterial causation. Philos. Rev. **33**(4), 331–333 (2013)
Sharpanskykh, O., Treur, J.: A temporal trace language for formal modelling and analysis of agent systems. In: Dastani, M., Hindriks, K.V., Meyer, J.-J.C. (eds.) Specification and verification of multi-agent systems, pp 317–353. Springer Verlag (2010)
Tinklepaugh, O.L.: Multiple delayed reaction with chimpanzees and monkeys. J. Comp. Psychol. **13**, 207–243 (1932)
Tollefsen, D.P.: From extended mind to collective mind. Cogn Systems Res **7**, 140–150 (2006)
Treur, J.: Temporal factorisation: a unifying principle for dynamics of the world and of mental states. Cogn. Syst. Res. **8**(2), 57–74 (2007a)
Treur, J.: Temporal factorisation: realisation of mediating state properties for dynamics. Cogn. Syst. Res. **8**(2), 75–88 (2007b)
Treur, J.: Past-future separation and normal forms in temporal predicate logic specifications. J. Algorithms Cogn. Inf. Logic **64**, 106–124 (2009)
Treur, J.: Network-Oriented Modeling: Addressing Complexity of Cognitive Affective and Social Interactions. Springer Cham (2016)
Treur, J.: Multilevel network reification: representing higher order adaptivity in a network. In: Aiello, L., Cherifi, C., Cherifi, H., Lambiotte, R., Lió, P., Rocha, L. (eds.) Complex Networks and Their Applications VII Proc ComplexNetworks'18 vol 1, Studies in Computational Intelligence, vol. 812, pp. 635–651, Springer Heidelberg (2018)
Treur, J.: The ins and outs of network-oriented modeling: from biological networks and mental networks to social networks and beyond. Trans. Comp. Coll. Intell. **32**, 120–139 (2019a)
Treur, J.: Adaptive network modeling for criterial causation. In: Cherifi, H., Gaito, S., Mendes, J.F., Moro, E., & Rocha, L.M. (eds.) Complex Networks and Their Applications VIII: Proceedings of the Eighth International Conference on Complex Networks and Their Applications Complex Networks 2019, vol. 2, Studies in Computational Intelligence, vol. 882, pp. 827–841. Springer Publishers (2019b)
Treur, J.: Network-Oriented Modeling for Adaptive Networks: Designing Higher-Order Adaptive Biological Mental and Social Network Models. Springer Nature Cham (2020a)
Treur, J.: Modeling higher-order adaptivity of a network by multilevel network reification. Netw. Sci. **8**(S1), S110–S144 (2020b)
Treur, J.: Modeling the emergence of informational content by adaptive networks for temporal factorisation and criterial causation. Cogn. Syst. Res. **68**, 34–52 (2021)
Tse, P.U.: The Neural Basis of Free Will: Criterial Causation. MIT Press, Cambridge (2013)
Tse, P.U.: Two types of libertarian free will are realized in the human brain. In: G.D. Caruso, O.J. Flanagan (eds) Neuroexistentialism: Meaning Morals and Purpose in the Age of Neuroscience, pp 248–290. Oxford University Press (2018)
van Gelder, T.J., Port, R.F.: It's About Time: An Overview of the Dynamical Approach to Cognition In: Port, RF van Gelder, T (eds) (1995), Mind as Motion: Explorations in the Dynamics of Cognition, pp 1–43. MIT Press Cambridge Mass (1995)