

# Reduction Relations for Agent Models

Jan Treur

Vrije Universiteit Amsterdam, Department of Artificial Intelligence  
De Boelelaan 1081, 1081 HV Amsterdam, The Netherlands  
treur@cs.vu.nl    <http://www.cs.vu.nl/~treur>

## Abstract

*This paper focuses on relationships between agent models and their physical realisations. Approaches on reduction from philosophical literature are analysed in a formalised manner, and extended by incorporating context-dependency w.r.t. specific makeups for their realisations. It is shown how these context-dependent reduction approaches can be translated into each other and how they can be applied to relate agent models.*

## 1. Introduction

Agent models can be designed at different levels of abstraction. For example, the well-known BDI-model (e.g., [21]) makes use of higher-level cognitive concepts such as beliefs, desires and intentions. Agent models can also be defined on the basis of the world's dynamics, and described by concepts at lower levels, for example, physical, chemical, or neurological concepts; e.g., [5, 6]. In order to ground models for embodied agents in a physical, chemical or neurological context, often the focus is on their interaction as a coupled system with the environment; e.g., [1, 7, 8, 13, 22]. However, they can be related to physical reality in a still more fundamental manner when the model of their internal functioning is fully immersed in a model of the world's dynamics, and to this end concepts from a lower level are used in the model, or it is indicated how the concepts used in the model relate to such lower-level concepts. In this way cognition can be addressed by an artificial life like approach; e.g., [6, 13, 19, 23]. This allows to model other types of mind-matter interaction, such as an agent that takes drugs which change its internal functioning (e.g., antidepressiva that affect neuro-transmitter levels), electrostimulation therapies, or brain-computer interfacing (e.g., [13]).

It is an interesting challenge to explore how exactly a higher-level agent model can be immersed in a lower-level model of the world's dynamics. A related fundamental question is whether this can be done in exactly one way, or in multiple ways. For example, is cognitive functioning depending on a neurological realisation, or is a different realisation also possible? Within the philosophical literature reduction approaches

are proposed to relate higher-level and lower-level descriptions (theories). In this paper three of these reduction approaches are analysed on applicability: the bridge law approach [18], the functional approach [14, 15, 16], and the interpretation mapping approach [24].

The paper is organised as follows. After a brief introduction to to reduction in Section 2, Section 3 shows how the three reduction approaches can be refined to incorporate the notion of a specific makeup, thus obtaining context-dependent variants that allow multiple realisation. Here a context is defined as a specific makeup within a lower-level theory. In Section 4 in a further comparative analysis it is shown under which conditions and how the approaches can be related to each other by mutual translations. In Section 5 a practical case study illustrates the applicability of the different approaches to relate a higher-level agent models to lower-level models.

## 2. Reduction Approaches

Work on reduction can be found in a wide variety of publications in the philosophical literature; see, for example [2, 3, 4, 14, 15, 16, 18]. Reduction addresses relationships between descriptions of two different levels, usually indicated by a higher-level theory  $T_2$  (e.g., a cognitive theory) and a lower-level or base theory  $T_1$  (e.g., a neurological theory). A specific reduction approach provides a particular *reduction relation*: a way in which each higher-level property  $a$  (an expression in  $T_2$ ) can be related to a lower-level property  $b$  (an expression in  $T_1$ ), this  $b$  is often called a *realiser* for  $a$ . Reduction approaches differ in how these relations are defined. In the classical *bridge law reduction* approach, following [18] reduction relations are specified by (biconditional) bridge principles  $a \leftrightarrow b$  that relate the expressions  $a$  in the language of a higher-level theory  $T_2$  in a one-to-one manner to expressions  $b$  in the language of the lower-level theory  $T_1$ . As an alternative Kim [16], pp. 98-102 describes *functional reduction* based on function-alisation of a target property  $a$  in  $T_2$  in terms of its causal task  $C$  (specifying its causal relationships to other properties) and relating it to a state property  $b$  in  $T_1$  performing this causal task  $C$ . For functional reduction the reduction relations are

not required to be one-to-one, thus allowing multiple realisation. A third notion to define reduction relations is a (*relative*) *interpretation mappings* (e.g., [2], [9], pp. 201-263; [22], pp. 61-65; [24]). These approaches relate the two theories  $T_2$  and  $T_1$  based on a mapping  $\varphi$  from expressions  $a$  of  $T_2$  to expressions  $b$  of  $T_1$ , in the sense that  $b = \varphi(a)$ .

From the three approaches mentioned, functional reduction is able to handle multiple realisation, but in an implicit manner. The interpretation mapping approach can be extended when multiple mappings are taken into account, so that multiple realisation is covered in an explicit manner. Bridge law reduction is not able to handle multiple realisation. However, in [14], pp. 233-236, Kim briefly sketches how what he calls a *local* or *structure-restricted* form of bridge law reduction, can handle multiple realisation. His suggestion is to relativise bridge principles to  $S \rightarrow (a \leftrightarrow b)$ , by adding an extra parameter  $S$  indicating the context of a specific system or makeup of an organism. Below in Section 3, following [25], it is shown how a variation on this idea of context-dependent reduction can be worked out in more detail for each of the three reduction approaches considered. Thus refined variants are obtained making multiple realisation explicit by reference to the context-dependency of a specific realisation. It will turn out that systematic relationships between these three refined approaches exist (see Section 4 for mutual translations).

### 3. Context-Dependent Reduction

In context-dependent reduction as introduced in [25], the aim is to identify a set of contexts and to relate the different realisations to these contexts. When contexts are defined in a sufficiently fine-grained manner, within one context the realisation can be unique. In this case, from an abstract viewpoint contexts can be seen as a form of parameterisation of the different possible realisations. For example, in Cognitive Science such a grouping could be based on species, i.e., groups of organisms with (more or less) the same makeup. If mental state properties (for example, having a certain sensory representation) are assumed that can be shared between, for example, biological organisms and robot-like systems it may be useful to allow contexts that are described within different base theories. Therefore in the context-dependent reduction approach developed, a collection of (base) theories  $\mathcal{T}_1$  is assumed and for each theory  $T$  in  $\mathcal{T}_1$  a set of contexts  $\mathcal{C}_T$ , such that each particular context<sup>1</sup> of an organism or system is formally described by a pair  $(T, S)$  of a specific theory  $T$  in  $\mathcal{T}_1$

<sup>1</sup> For the sake of shortness a context  $(T, S)$  is often indicated just by  $S$ .

together with a specific context  $S$  in  $\mathcal{C}_T$ . The contexts  $S$  are assumed to be descriptions in the language of  $T$  and consistent with  $T$ . The theories  $T$  in  $\mathcal{T}_1$  and contexts  $S$  in  $\mathcal{C}_T$  can be used to distinguish the different realisations that are possible. Below it is shown how this can be done for the three approaches considered. Note that when the collection of theories  $\mathcal{T}_1$  is taken a singleton  $\{T_1\}$  consisting of one theory  $T_1$  and the set of contexts  $\mathcal{C}_{T_1}$  is taken a singleton  $\{S\}$  consisting of the empty specification  $S = \phi$ , then the original general reduction approach is obtained.

#### Context-dependent bridge law reduction

For the bridge law reduction approach, the set of realisers that exists within one context  $S$  for a theory  $T$  in  $\mathcal{T}_1$ , is expressed by context-dependent biconditional bridge laws parameterised by a theory  $T$  in  $\mathcal{T}_1$  and a context  $S$  in  $\mathcal{C}_T$ , specified by

$$a_1 \leftrightarrow b_{1,T,S}, \dots, a_k \leftrightarrow b_{k,T,S}$$

Given such a parameterised specification, the context-dependent criterion of bridge law reduction for a law  $L(a_1, \dots, a_k)$  derivable from  $T_2$  can be formulated (in two equivalent manners) by<sup>2</sup>:

- (i)  $T_2 \vdash L(a_1, \dots, a_k) \Rightarrow$   
 $\forall T \in \mathcal{T}_1 \forall S \in \mathcal{C}_T$   
 $T \cup S \cup \{a_1 \leftrightarrow b_{1,T,S}, \dots, a_k \leftrightarrow b_{k,T,S}\} \vdash L(a_1, \dots, a_k)$
- (ii)  $T_2 \vdash L(a_1, \dots, a_k) \Rightarrow$   
 $\forall T \in \mathcal{T}_1 \forall S \in \mathcal{C}_T T \cup S \vdash L(b_{1,T,S}, \dots, b_{k,T,S})$

Note that context-dependent bridge law reduction implies unique realisers (up to equivalence) per context: from  $a \leftrightarrow b_{T,S}$  and  $a \leftrightarrow b'_{T,S}$  it follows that  $b_{T,S}$  and  $b'_{T,S}$  cannot be non-equivalent in  $T \cup S$ . So to obtain context-dependent bridge law reduction in cases of multiple realisation, the contexts are defined with a grain-size such that per context a unique realisation exists.

#### Context-dependent functional reduction

For a given collection of context theories  $\mathcal{T}_1$  and sets of contexts  $\mathcal{C}_T$ , for context-dependent functional reduction a first criterion is that a joint causal role specification<sup>3</sup>  $C(P_1, \dots, P_k)$  can be identified such that it covers all relevant state properties of theory  $T_2$ , and for each

<sup>2</sup> Note that the notation  $L(a_1, \dots, a_k)$  is used to indicate how a more complex statement is built as a proposition from subformulae  $a_1, \dots, a_k$ . Furthermore, the theory  $T \cup S \cup \{a_1 \leftrightarrow b_{1,T,S}, \dots, a_k \leftrightarrow b_{k,T,S}\}$  is a *definitional extension* of  $T \cup S$  obtained by adding (new) symbols for  $a_i$  to  $T$ ; see, for example [9], p. 60; [22], p. 57-61. Every definitional extension is a *conservative extension*; therefore for all statements  $\alpha$  in the language of  $T$  it holds  $T \cup S \cup \{a_1 \leftrightarrow b_{1,T,S}, \dots, a_k \leftrightarrow b_{k,T,S}\} \vdash \alpha$  if and only if  $T \cup S \vdash \alpha$ ; see, e.g., [9], pp. 59-60, 66; [22], p. 41, 57-61.

<sup>3</sup> This specifies the causal relationships of these properties to each other and to other (external) properties; this can be obtained by the Ramsey-Lewis method as described in [14], [17], [20].

theory  $T$  in  $\mathfrak{T}_1$  and context  $S$  in  $\mathbf{C}_T$  at least one instantiation of it within  $T$  exists:

$$\forall T \in \mathfrak{T}_1 \forall S \in \mathbf{C}_T \exists P_1, \dots, P_k \quad T \cup S \vdash C(P_1, \dots, P_k).$$

The second criterion for context-dependent functional reduction, concerning laws is

$$T_2 \vdash L(a_1, \dots, a_k) \Rightarrow \forall T \in \mathfrak{T}_1 \forall S \in \mathbf{C}_T \forall P_1, \dots, P_k \\ [T \cup S \vdash C(P_1, \dots, P_k) \Rightarrow T \cup S \vdash L(P_1, \dots, P_k)]$$

In general this notion of context-dependent functional reduction may still allow multiple realisation within one theory and context. However, by choosing contexts with an appropriate grain-size it can be achieved that within one given theory and context unique realisation occurs. The *unique realisation context criterion* (also called *strictness criterion*) expresses this as follows. For each  $T$  in  $\mathfrak{T}_1$  and context  $S$  in  $\mathbf{C}_T$  there exists a unique set of instantiations realising the joint causal role specification  $C(P_1, \dots, P_k)$ , or formally:

$$\forall T \in \mathfrak{T}_1 \forall S \in \mathbf{C}_T \exists P_1, \dots, P_k \\ [T \cup S \vdash C(P_1, \dots, P_k) \ \& \ \\ \forall Q_1, \dots, Q_k \ [T \cup S \vdash C(Q_1, \dots, Q_k) \Rightarrow \\ T \cup S \vdash P_1 \leftrightarrow Q_1 \ \& \dots \ \& \ P_k \leftrightarrow Q_k]]$$

This guarantees per theory  $T$  and context  $S$  unique realisers, parameterised by  $T$  and  $S$ . When also this third criterion is satisfied, a form of reduction is obtained that we call *strict context-dependent functional reduction*. When the strictness criterion is satisfied, the universally quantified form for relations between laws is equivalent to an existentially quantified variant:

$$T_2 \vdash L(a_1, \dots, a_k) \Rightarrow \forall T \in \mathfrak{T}_1 \forall S \in \mathbf{C}_T \exists P_1, \dots, P_k \\ [T \cup S \vdash C(P_1, \dots, P_k) \ \& \ T \cup S \vdash L(P_1, \dots, P_k)]$$

### Context-dependent interpretation

To obtain a form of context-dependent interpretation, the notion of interpretation mapping is generalised to a multi-mapping, parameterised by contexts. A *context-dependent interpretation* of a theory  $T_2$  in a collection of theories  $\mathfrak{T}_1$  with sets of contexts  $\mathbf{C}_T$  specifies for each theory  $T$  in  $\mathfrak{T}_1$  and context  $S$  in  $\mathbf{C}_T$  an appropriate mapping  $\varphi_{T,S}$  from the expressions of  $T_2$  to expressions of  $T$ : a multi-mapping

$$\varphi_{T,S} (T \in \mathfrak{T}_1, S \in \mathbf{C}_T)$$

from theory  $T_2$  to theories  $T$  in  $\mathfrak{T}_1$  parameterised by theories  $T$  in  $\mathfrak{T}_1$  and contexts  $S$  in  $\mathbf{C}_T$ . Such a multi-mapping is a context-dependent interpretation mapping when it satisfies the property that if a law  $L$  can be derived from  $T_2$ , then for each  $T$  in  $\mathfrak{T}_1$  and context  $S$  in  $\mathbf{C}_T$  the statement  $\varphi_{T,S}(L)$  can be derived from  $T \cup S$ :

$$T_2 \vdash L \Rightarrow \forall T \in \mathfrak{T}_1 \forall S \in \mathbf{C}_T \quad T \cup S \vdash \varphi_{T,S}(L)$$

Usually the mappings are assumed compositional with respect to logical connectives<sup>4</sup>. Note that also here within one theory  $T$  in  $\mathfrak{T}_1$  and context  $S$  in  $\mathbf{C}_T$  multiple realisation is possible, expressed as the existence of two essentially different interpretation mappings  $\varphi_{T,S}$  and  $\varphi'_{T,S}$ , i.e. such that in  $T \cup S$  it may not hold that  $\varphi_{T,S}(a) \leftrightarrow \varphi'_{T,S}(a)$ . However, an additional *strictness criterion* to obtain unique realisation per context is formulated as follows: when for any given theory  $T$  in  $\mathfrak{T}_1$  and context  $S$  in  $\mathbf{C}_T$  two interpretation mapping  $\varphi_{T,S}$  and  $\varphi'_{T,S}$  are possible, then for all  $a$  it holds that

$$T \cup S \vdash \varphi_{T,S}(a) \leftrightarrow \varphi'_{T,S}(a)$$

When this additional criterion is satisfied as well, the interpretation is called a *strict context-dependent interpretation*.

## 4. Mutual Translations

In [25] the question in how far the different approaches to context-dependent reduction can be related to each other was not addressed. In this section it is shown how the context-dependent interpretation mapping approach can be related to the other two context-dependent approaches by mutual translations.

### Relating bridge law reduction interpretation

In this subsection it is shown how bridge law reduction can be translated into reduction based on a strict interpretation mapping and vice versa.

#### From interpretation to bridge law reduction

Suppose a strict interpretation mapping  $\varphi_{T,S}$  is given, which is assumed compositional. For each basic expression  $a$  of  $T_2$  specify the bridge principle

$$a_i \leftrightarrow b_{i,T,S} \quad \text{with } b_{i,T,S} = \varphi_{T,S}(a_i)$$

If  $L(a_1, \dots, a_k)$  is law derivable from  $T_2$  involving state properties  $a_1, \dots, a_k$ , then

$$T \cup S \vdash \varphi_{T,S}(L(a_1, \dots, a_k)).$$

By compositionality of  $\varphi$  it follows that

$$T \cup S \vdash L(\varphi_{T,S}(a_1), \dots, \varphi_{T,S}(a_k)).$$

Therefore it follows

$$T \cup S \vdash L(b_{1,T,S}, \dots, b_{k,T,S}).$$

This shows that the criterion (ii) for bridge law reduction is fulfilled. Note that it is needed to assume the interpretation to be strict. If the same translation would be done for two essentially different non-strict interpretations  $\varphi_{T,S}$  and  $\varphi'_{T,S}$ , it would lead to contradictions.

#### From bridge law reduction to interpretation

For a translation the other way around, assume bridge principles

<sup>4</sup> This means that  $\varphi_{T,S}(A1 \ \& \ A2) = \varphi_{T,S}(A1) \ \& \ \varphi_{T,S}(A2)$ , and so on.

$$a_i \leftrightarrow b_{i,T,S}$$

are given for the basic expressions  $a_i$  of  $T_2$ , such that the bridge law reduction criterion (ii) is fulfilled:

$$T_2 \vdash L(a_1, \dots, a_k) \Rightarrow T \cup S \vdash L(b_{1,T,S}, \dots, b_{k,T,S})$$

Define the mapping  $\varphi_{T,S}$  as follows. For each basic expression  $a_i$  of  $T_2$ , based on the given bridge principle  $a_i \leftrightarrow b_{i,T,S}$ , define

$$\varphi_{T,S}(a_i) = b_{i,T,S}$$

For more complex expressions extend this by compositionality:

$$\begin{aligned} \varphi_{T,S}(A_1 \wedge A_2) &= \varphi_{T,S}(A_1) \wedge \varphi_{T,S}(A_2) \\ \varphi_{T,S}(A_1 \vee A_2) &= \varphi_{T,S}(A_1) \vee \varphi_{T,S}(A_2) \\ \varphi_{T,S}(A_1 \rightarrow A_2) &= \varphi_{T,S}(A_1) \rightarrow \varphi_{T,S}(A_2) \\ \varphi_{T,S}(\neg A) &= \neg \varphi_{T,S}(A) \\ \varphi_{T,S}(\forall x A(x)) &= \forall x \varphi_{T,S}(A(x)) \\ \varphi_{T,S}(\exists x A(x)) &= \exists x \varphi_{T,S}(A(x)) \end{aligned}$$

For this mapping  $\varphi_{T,S}$ , from  $T_2 \vdash L(a_1, \dots, a_k)$  by the bridge law reduction criterion (ii) it follows by compositionality:

$$\begin{aligned} T_2 \vdash L(a_1, \dots, a_k) &\Rightarrow \\ T \cup S \vdash L(\varphi_{T,S}(a_1), \dots, \varphi_{T,S}(a_k)) &\Rightarrow \\ T \cup S \vdash \varphi_{T,S}(L(a_1, \dots, a_k)). & \end{aligned}$$

Therefore the criterion for an interpretation mapping is fulfilled.

Note that the two translations from bridge law reduction to interpretation and from interpretation to bridge law reduction as given are each others' inverse. Moreover, note that the context-dependent interpretation obtained from bridge law reduction is strict. When within a given theory and context an essentially different interpretation would be possible, this could be translated into bridge laws as well, which would lead to a contradiction.

### Relating interpretation to functional reduction

Next it is shown how functional reduction can be translated into an interpretation and vice versa.

#### From functional reduction to interpretation

Let a theory  $T$  in  $\mathcal{T}_I$  and a context  $S$  in  $\mathbf{C}_T$  be given. Take a joint causal role specification  $C(P_1, \dots, P_k)$  for the basic state properties  $a_1, \dots, a_k$  of  $T_2$ . Suppose  $L(a_1, \dots, a_k)$  is derivable from  $T_2$  is given and the functional reduction criterion holds:

$$\begin{aligned} T_2 \vdash L(a_1, \dots, a_k) &\Rightarrow \forall P_1, \dots, P_k \\ [T \cup S \vdash C(P_1, \dots, P_k) &\Rightarrow T \cup S \vdash L(P_1, \dots, P_k)] \end{aligned}$$

Pick an arbitrary set  $b_1, \dots, b_k$  of realisers satisfying  $C(P_1, \dots, P_k)$  in  $T \cup S$ , and define

$$\varphi_{T,S}(a_i) = b_i$$

For more complex expressions extend this by compositionality as above. For  $L(a_1, \dots, a_k)$  derivable from  $T_2$ , by the functional reduction criterion it holds:

$$\begin{aligned} T \cup S \vdash C(\varphi_{T,S}(a_1), \dots, \varphi_{T,S}(a_k)) &\Rightarrow \\ T \cup S \vdash L(\varphi_{T,S}(a_1), \dots, \varphi_{T,S}(a_k)) & \end{aligned}$$

As the antecedent holds due to the choice of the mapping, it follows that

$$T \cup S \vdash L(\varphi_{T,S}(a_1), \dots, \varphi_{T,S}(a_k)).$$

By the compositional definition of  $\varphi_{T,S}$  for more complex expressions, as before, it follows:

$$T \cup S \vdash \varphi_{T,S}(L(a_1, \dots, a_k)).$$

Therefore the interpretation mapping criterion is fulfilled for the chosen mapping  $\varphi_{T,S}$ . Note that a mapping  $\varphi_{T,S}$  as defined above fully depends on the chosen set of realisers of the joint causal role specification  $C(P_1, \dots, P_k)$ . This may result in a collection of possible mappings for each instantiation of  $P_1, \dots, P_k$  satisfying  $C(P_1, \dots, P_k)$  in  $T \cup S$ . This can be avoided by assuming the additional criterion of unique realisation context:

$$\begin{aligned} \exists P_1, \dots, P_k [T \cup S \vdash C(P_1, \dots, P_k) \ \& \ \forall Q_1, \dots, Q_k \\ [T \cup S \vdash C(Q_1, \dots, Q_k) \Rightarrow \\ T \cup S \vdash P_1 \leftrightarrow Q_1 \ \& \ \dots \ \& \ P_k \leftrightarrow Q_k]] \end{aligned}$$

Using this (i.e., assuming strict context-dependent functional reduction), per theory  $T$  and context  $S$  a unique interpretation mapping is found. This is a strict interpretation mapping, because any essentially different interpretation mapping would provide another set of realisers within the same context.

#### From interpretation to functional reduction

Suppose a context-dependent interpretation mapping  $\varphi_{T,S}$  ( $T \in \mathcal{T}_I$ ,  $S \in \mathbf{C}_T$ ) is given, which is assumed compositional. Moreover, let  $L(a_1, \dots, a_k)$  be derivable from  $T_2$ . Let a theory  $T$  in  $\mathcal{T}_I$  and a context  $S$  in  $\mathbf{C}_T$  be given. Then by the interpretation mapping criterion it holds  $T \cup S \vdash \varphi_{T,S}(L(a_1, \dots, a_k))$  and hence by the compositionality assumption it holds

$$T \cup S \vdash L(\varphi_{T,S}(a_1), \dots, \varphi_{T,S}(a_k)).$$

Assume that a joint causal role specification is given by  $C(a_1, \dots, a_k)$ , which means  $T_2 \vdash C(a_1, \dots, a_k)$  holds. The interpretation mapping criterion applied to  $C(a_1, \dots, a_k)$  provides

$$T_2 \vdash C(a_1, \dots, a_k) \Rightarrow T \cup S \vdash \varphi_{T,S}(C(a_1, \dots, a_k)).$$

By the compositionality assumption it holds

$$\varphi_{T,S}(C(a_1, \dots, a_k)) = C(\varphi_{T,S}(a_1), \dots, \varphi_{T,S}(a_k)).$$

Hence from  $T_2 \vdash C(a_1, \dots, a_k)$  it follows

$$T \cup S \vdash C(\varphi_{T,S}(a_1), \dots, \varphi_{T,S}(a_k)).$$

Therefore, if the variables  $P_1, \dots, P_k$  are instantiated by  $\varphi_{T,S}(a_1), \dots, \varphi_{T,S}(a_k)$ , it holds

$$T \cup S \vdash C(P_1, \dots, P_k) \ \& \ T \cup S \vdash L(P_1, \dots, P_k)$$

Now this only shows that for some instantiation of the variables  $P_1, \dots, P_k$  the functional reduction criterion holds, not for all instantiations. In fact for the general case the following weaker existential criterion is implied:

$$T_2 \vdash L(a_1, \dots, a_k) \Rightarrow \\ \exists P_1, \dots, P_k [T \cup S \vdash C(P_1, \dots, P_k) \& \\ T \cup S \vdash L(P_1, \dots, P_k)]$$

This weaker existential criterion is (only) equivalent to the stronger universal criterion when it is assumed that exactly one unique set of instantiations of the variables  $P_1, \dots, P_k$  is possible within context  $S$  for theory  $T$  such that  $C(P_1, \dots, P_k)$  holds (unique realisation criterion). Therefore to obtain a faithful translation it is assumed that  $\varphi_{T,S}$  is a strict context-dependent interpretation. In such a case if an essentially different realising instantiation of  $C(P_1, \dots, P_k)$  would be possible, this would lead to an essentially different interpretation mapping (see previous translation), which is excluded by the criterion of strictness. Thus by the translation discussed above, both (equivalent) universal and existential versions of the criterion for context-dependent functional reduction are satisfied.

## 5. Relating Agent Models

In this section the concepts discussed above are illustrated for a case study involving a higher-level cognitive model  $CM$  and two lower-level models: a neurological model  $NM$  and a biochemical model  $BM$ . The neurological model  $NM$  will be described by a general neurological theory  $NT$  and a specific makeup  $NS$ , describing a specific context. Similarly, the biochemical model  $BM$  is described by a general biochemical theory  $CT$  and a specific makeup  $CS$ , which describes (in a simplified form) the context of the bacterium *E. coli*. Based on the neurological theory  $NT$  and biochemical theory  $CM$  and the neural example context  $NS$  and the biochemical context  $BS$  of *E. coli*, the context-dependent interpretation  $(\varphi_{NT,NS}, \varphi_{BT,BS})$  for the cognitive model  $CM$  is defined:

$$T_2 = CM, \mathcal{T}_1 = \{NT, CT\}, \mathbf{C}_{NT} = \{NS\}, \mathbf{C}_{CT} = \{CS\}.$$

**Cognitive model  $CM$**  This model plays the role of the higher-level theory. It describes a simple cognitive process which depending on observations on two world facts  $s1$  and  $s2$  makes a choice between two actions  $a1$  and  $a2$ . It is specified as follows:

$$\begin{aligned} worldfact(X) &\rightarrow observed(X) \\ observed(X) &\rightarrow belief(X) \\ belief(s1) \& \text{ not } belief(s2) &\rightarrow intention(a1) \\ belief(s2) &\rightarrow intention(a2) \end{aligned}$$

$$\begin{aligned} intention(a1) \& \text{ belief}(s1) &\rightarrow performed(a1) \\ intention(a2) \& \text{ belief}(s2) &\rightarrow performed(a2) \\ performed(a1) \& \text{ worldfact}(s1) &\rightarrow worldfact(e1) \\ performed(a2) \& \text{ worldfact}(s2) &\rightarrow worldfact(e2) \end{aligned}$$

**Neurological model  $NM$**  For the neurological model a situation is taken involving two objects. When a cube is seen and no sphere, it will be taken, when a sphere is seen and no cube, it will be taken. When both are seen, only the sphere is taken. The neurological model  $NM$  used consists of the general laws specified in (simplified) neurological theory  $NT$  and a specific neural makeup described by  $NS$ .

**Neurological theory  $NT$**  Activations of neurons propagate via connections through synapses with positive (excitatory) or negative (inhibitory) effects. In case of multiple connections to one neuron, the effect is combined, and activation takes place when this combined input is above the neuron's threshold. When a sensor stimulus occurs that is connected to a neuron, then this neuron is activated, when the input is above its threshold. When the combined input for an action is above its threshold, then this action occurs. This is formalised as:

$$\begin{aligned} connectedto(X, Y, pos) \& \text{ activated}(X) \& \\ \text{ threshold}(Y, v) \& v < 1 &\rightarrow \text{ activated}(Y) \\ connectedto(X1, X2, Y, pos) \& \text{ activated}(X1) \& \text{ activated}(X2) \\ \& \text{ threshold}(Y, v) \& v < 2 &\rightarrow \text{ activated}(Y) \\ connectedto(X1, Y, pos) \& \text{ connectedto}(X2, Y, neg) \& \\ \text{ activated}(X1) \& \text{ not activated}(X2) \& \\ \text{ threshold}(Y, v) \& v < 1 &\rightarrow \text{ activated}(Y) \\ occurs(X) &\rightarrow seeing(X) \\ activated(\text{take}(X)) &\rightarrow having(X) \end{aligned}$$

Note that here the predicate *connectedto* is used to represent all combined positive input for a neuron and all combined negative input for a neuron. Moreover, note that for convenience in the last line some world relations have been included.

**Neural makeup  $NS$**  (see Fig. 1):

$$\begin{aligned} connectedto(\text{seeing}(\text{cube}), SN1, pos) \\ connectedto(\text{seeing}(\text{sphere}), SN2, pos) \\ connectedto(SN2, MN2, \text{take}(\text{sphere}), pos) \\ connectedto(SN1, MN1, \text{take}(\text{cube}), pos) \\ connectedto(SN1, MN1, pos) \\ connectedto(SN2, MN2, pos) \\ connectedto(SN2, MN1, neg) \\ \text{ threshold}(SN1, 0.5) \\ \text{ threshold}(SN2, 0.5) \\ \text{ threshold}(MN1, 0.5) \\ \text{ threshold}(MN2, 0.5) \\ \text{ threshold}(\text{take}(\text{cube}), 1.5) \\ \text{ threshold}(\text{take}(\text{sphere}), 1.5) \end{aligned}$$

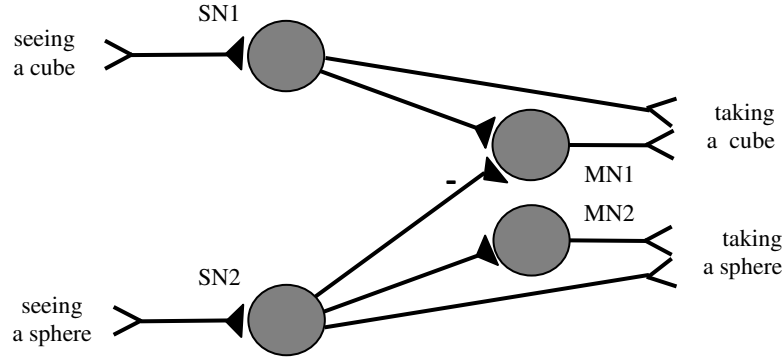


Figure 1 Neural makeup NS

### Mapping the cognitive model onto the neural model

Given the cognitive model  $CM$  and the neurological model  $NM$ , the next step is to relate them by a reduction relation. As in Section 4 it was shown how the different context-dependent reduction approaches can be translated into each other, it is only shown for one of them: the interpretation mapping approach. The cognitive model  $CM$  is mapped onto the neurological model  $NM$  by the interpretation mapping  $\varphi_{NT,NS}$  defined by (and extended to more complex propositions in a compositional manner):

$$\begin{aligned} \varphi_{NT,NS}(\text{observed}(s1)) &= \text{seeing}(\text{cube}) \\ \varphi_{NT,NS}(\text{observed}(s2)) &= \text{seeing}(\text{sphere}) \\ \varphi_{NT,NS}(\text{belief}(s1)) &= \text{activated}(SN1) \\ \varphi_{NT,NS}(\text{belief}(s2)) &= \text{activated}(SN2) \\ \varphi_{NT,NS}(\text{intention}(a1)) &= \text{activated}(MN1) \\ \varphi_{NT,NS}(\text{intention}(a2)) &= \text{activated}(MN2) \\ \varphi_{NT,NS}(\text{performed}(a1)) &= \text{activated}(\text{take}(\text{cube})) \\ \varphi_{NT,NS}(\text{performed}(a2)) &= \text{activated}(\text{take}(\text{sphere})) \end{aligned}$$

For example, the relation

$$\text{belief}(s1) \ \& \ \text{not} \ \text{belief}(s2) \rightarrow \text{intention}(a1)$$

of the cognitive model is mapped by  $\varphi_{NT,NS}$  as follows:

$$\begin{aligned} \varphi_{NT,NS}(\text{belief}(s1) \ \& \ \text{not} \ \text{belief}(s2) \rightarrow \text{intention}(a1)) \\ &= \varphi_{NT,NS}(\text{belief}(s1) \ \& \ \text{not} \ \text{belief}(s2)) \rightarrow \varphi_{NT,NS}(\text{intention}(a1)) \\ &= \varphi_{NT,NS}(\text{belief}(s1)) \ \& \ \varphi_{NT,NS}(\text{not} \ \text{belief}(s2)) \rightarrow \\ &\quad \varphi_{NT,NS}(\text{intention}(a1)) \\ &= \varphi_{NT,NS}(\text{belief}(s1)) \ \& \ \text{not} \ \varphi_{NT,NS}(\text{belief}(s2)) \rightarrow \\ &\quad \varphi_{NT,NS}(\text{intention}(a1)) \\ &= \text{activated}(SN1) \ \& \ \text{not} \ \text{activated}(SN2) \rightarrow \text{activated}(MN1) \end{aligned}$$

From  $NT \cup NS$  the following relationships can be derived:

$$\begin{aligned} \text{occurs}(\text{cube}) &\rightarrow \text{seeing}(\text{cube}) \\ \text{occurs}(\text{sphere}) &\rightarrow \text{seeing}(\text{sphere}) \\ \text{seeing}(\text{cube}) &\rightarrow \text{activated}(SN1) \\ \text{seeing}(\text{sphere}) &\rightarrow \text{activated}(SN2) \\ \text{activated}(SN2) &\rightarrow \text{activated}(MN2) \\ \text{activated}(SN1) \ \& \ \text{not} \ \text{activated}(SN2) &\rightarrow \text{activated}(MN1) \end{aligned}$$

$$\begin{aligned} \text{activated}(MN1) \ \& \ \text{activated}(SN1) &\rightarrow \\ &\quad \text{activated}(\text{take}(\text{cube})) \\ \text{activated}(MN2) \ \& \ \text{activated}(SN1) &\rightarrow \\ &\quad \text{activated}(\text{take}(\text{sphere})) \\ \text{activated}(\text{take}(\text{cube})) \ \& \ \text{occurs}(\text{cube}) &\rightarrow \\ &\quad \text{having}(\text{cube}) \\ \text{activated}(\text{take}(\text{sphere})) \ \& \ \text{occurs}(\text{sphere}) &\rightarrow \\ &\quad \text{having}(\text{sphere}) \end{aligned}$$

These relationships are exactly the mapped relationships from  $CM$ , formally:  $\varphi_{NT,NS}(CM)$ . This shows that the criterion for interpretation is satisfied.

### Biochemical model BM

Within cell biology causal chains are known in the form of chemical pathways from the environment to within the cell. For example, such causal chains justify to interpret the presence of an internal concentration of CRPcAMP above a certain level as an indicator for 'glucose being absent in the external environment', and of the internal presence of a certain concentration of lactose as an indicator for 'lactose being present in the external environment'. This shows ways in which a cell is able to build and maintain internal states that can be interpreted as a world model, or its beliefs about the world. Intentions can be considered to be present in the cell in that, depending on the observed environment it is able to make an informed choice (for preparation of an action) between alternatives of specific import action to provide resources for a specific type of metabolism. See [10, 11] for more detailed models for intracellular processes underlying bacterial behaviour. For the simplified example considered here, the biochemical theory  $BT$  consists of general biochemical laws indicating how certain types of substances in general can react with each other. The makeup  $BS$  of  $E.coli$  specifies the presence of a specific cell membrane, a water-like fluid inside of appropriate temperature, and the presence of specific substances within the cell, such as DNA.

## Mapping the cognitive onto the biochemical model

The cognitive model  $CM$  can be mapped onto the biochemical model  $BM$  by the interpretation mapping  $\varphi_{BT,BS}$  defined by:

$$\begin{aligned}\varphi_{NT,NS}(\text{observation}(s1)) &= \text{external lactose} \\ \varphi_{NT,NS}(\text{observation}(s2)) &= \text{external glucose} \\ \varphi_{BT,BS}(\text{belief}(s1)) &= \text{some lactose} \\ \varphi_{BT,BS}(\text{belief}(s2)) &= \text{no CRPcAMP} \\ \varphi_{BT,BS}(\text{intention}(a1)) &= \text{lactose import enzyme} \\ \varphi_{BT,BS}(\text{intention}(a2)) &= \text{glucose import enzyme} \\ \varphi_{NT,NS}(\text{action}(a1)) &= \text{import lactose} \\ \varphi_{NT,NS}(\text{action}(a2)) &= \text{import glucose}\end{aligned}$$

Again the mapping is extended for more complex propositions in a compositional manner. For example, the relation

$$\text{belief}(s1) \ \& \ \text{not} \ \text{belief}(s2) \rightarrow \text{intention}(a1)$$

of the cognitive model is mapped by  $\varphi_{BT,BS}$  as follows:

$$\begin{aligned}\varphi_{BT,BS}(\text{belief}(s1) \ \& \ \text{not} \ \text{belief}(s2) \rightarrow \text{intention}(a1)) \\ &= \varphi_{BT,BS}(\text{belief}(s1) \ \& \ \text{not} \ \text{belief}(s2)) \rightarrow \varphi_{BT,BS}(\text{intention}(a1)) \\ &= \varphi_{BT,BS}(\text{belief}(s1)) \ \& \ \varphi_{BT,BS}(\text{not} \ \text{belief}(s2)) \rightarrow \\ &\quad \varphi_{BT,BS}(\text{intention}(a1)) \\ &= \varphi_{BT,BS}(\text{belief}(s1)) \ \& \ \text{not} \ \varphi_{BT,BS}(\text{belief}(s2)) \rightarrow \\ &\quad \varphi_{BT,BS}(\text{intention}(a1))\end{aligned}$$

$$\begin{aligned}&= \text{some lactose} \ \& \ \text{not} \ \text{no CRPcAMP} \rightarrow \\ &\quad \text{lactose import enzyme}\end{aligned}$$

From  $BT \cup BS$  the following relationships can be derived. For example, for sensory processes the following derivable relationships describe that the external presence of glucose and or lactose leads to the presence of the related internal indicators.

$$\begin{aligned}\text{external glucose} &\rightarrow \text{no CRPcAMP} \\ \text{external lactose} &\rightarrow \text{some lactose} \\ \text{no CRPcAMP} &\rightarrow \text{glucose import enzyme} \\ \text{glucose import enzyme} &\rightarrow \text{glucose import} \\ \text{lactose import enzyme} &\rightarrow \text{lactose import} \\ \text{some lactose} \ \& \ \text{not} \ \text{no CRPcAMP} &\rightarrow \text{lactose import enzyme}\end{aligned}$$

The relationships for action generation derivable from  $BT \cup BS$  cover transcription (DNA affecting the presence of mRNA), translation (mRNA affecting the presence of enzyme), catalysis (enzymes affecting the related import reactions), and the effects of (co)factors in these steps (also see Fig. 2). These relationships are the mapped relationships from  $CM$ , formally:  $\varphi_{BT,BS}(CM)$ . This illustrates the fulfilment of the criterion for interpretation.

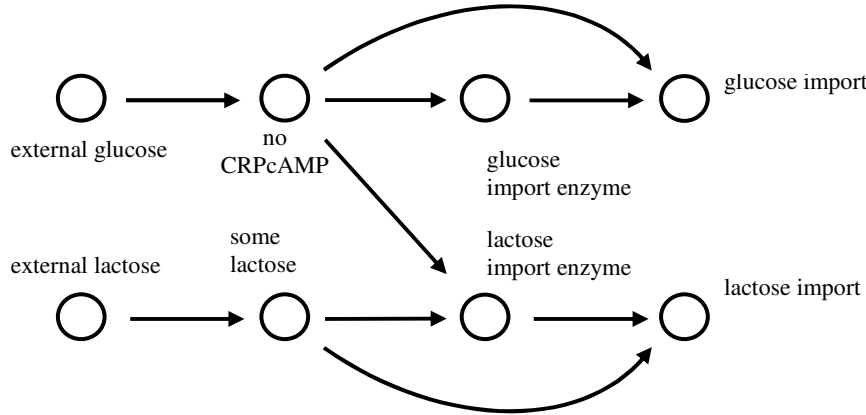


Figure 2 Derived biochemical relationships for *E. coli*

## 6. Discussion

Agents described by a higher-level model can have different physical realisations. In this paper it was shown how some of the approaches on reduction available in philosophical literature can be applied to relate such a higher-level agent model to its physical realisations. As these approaches do not treat multiple realisation in an explicit manner, refined variants of all three approaches (adopted from [25]) were used making multiple realisation explicit by reference to the context-dependency of a specific realisation. The notion of context-dependency distinguishes the general rules or

laws of an underlying theory from more specific aspects such as a particular makeup of an agent, for example, the general rules for neural systems in contrast to a particular neural architecture. It turned out to be possible to obtain systematic relationships between the three refined context-dependent reduction approaches, in the form of mutual translations between them (which was not addressed in [25]). The treatment as presented abstracts from the dynamic aspects (as is usually done in the philosophical literature mentioned). A topic for future work is to make these dynamic aspects explicit, by considering a form of temporal logic.

In a case study it was shown how based on the machinery developed an example cognitive agent model can be related both to a realisation by a neural model and by a biochemical model. Here the neural and the biochemical model each consist of a generic part specifying general laws or rules, and a specific part specifying the particular makeup considered. The higher-level cognitive model unifies basic properties of the two different lower-level models and describes them in a more abstract manner.

By having (in addition) a realisation of a higher-level agent model, it becomes possible to incorporate mutual effects between the physical world and an agent's internal functioning, for example, drugs which affect cognitive functioning, or brain-computer interfacing, as described in [13]. Within the framework presented here, effects of the world on the agent's functioning in principle can be modelled as a change from the agent's makeup  $S$  to a changed makeup  $S'$  (for example, a drug that affects the activation threshold of neurons, or inactivation of certain genes at the cell's DNA). In general such a changed makeup  $S'$  can be a realisation of the same or of another higher-level agent model.

## References

- [1] Bickhard, M.H. (1993). Representational Content in Humans and Machines. *J. of Experimental and Theoretical AI*, vol. 5, 1993, pp. 285-333.
- [2] Bickle, J. (1992). Mental Anatomy and the New Mind-Brain Reductionism. *Philosophy of Science*, vol. 59, pp. 217-230.
- [3] Bickle, J. (1998). *Psychoneural Reduction: The New Wave*. MIT Press, Cambridge, Mass.
- [4] Bickle, J. (2003). *Philosophy and Neuroscience*. Kluwer Academic Publishers.
- [5] Bosse, T., Jonker, C.M., and Treur, J., (2007). Simulation and Analysis of Adaptive Agents: an Integrative Modelling Approach. *Advances in Complex Systems*, vol. 10, 2007, pp. 335 - 357.
- [6] Bosse, T., and Treur, J., Formalising Agency-Inducing Patterns in World Dynamics. In: Rocha, L.M., et al. (eds.), *Artificial Life X: Proc. of the 10th International Conference*. MIT Press, 2006, pp. 546-552.
- [7] Clancey, W. (1997). *Situated Cognition: On Human Knowledge and Computer Representations*. Cambridge University Press.
- [8] Clark, A. (1997). *Being There: Putting Brain Body and World Together Again*. Cambridge, MA: MIT Press.
- [9] Hodges, W. (1993). *Model theory*. Cambridge University Press.
- [10] Jonker, C.M., Snoep, J.L., Treur, J., Westerhoff, H.V., and Wijngaards, W.C.A. (2002). Putting Intentions into Cell Biochemistry: An Artificial Intelligence Perspective. *Journal of Theoretical Biology*, vol. 214, pp. 105-134.
- [11] Jonker, C.M., Snoep, J.L., Treur, J., Westerhoff, H.V., Wijngaards, W.C.A., (2008). BDI-Modelling of Complex Intracellular Dynamics. *Journal of Theoretical Biology*, vol. 251, 2008, pp. 1-23.
- [12] Jonker, C.M., and Treur, J., A Temporal-Interactivist Perspective on the Dynamics of Mental States. *Cognitive Systems Research*, vol. 4, 2003, pp. 137-155.
- [13] Jonker, C.M., and Treur, J., Modelling Multiple Mind-Matter Interaction. *International Journal of Human-Computer Studies*, vol. 57, 2002, pp. 165-214.
- [14] Kim, J. (1996). *Philosophy of Mind*. Westview Press.
- [15] Kim, J. (1998). *Mind in a Physical world*. MIT Press.
- [16] Kim, J. (2005). *Physicalism, or Something Near Enough*. Princeton University Press, Princeton.
- [17] Lewis, D.K. (1972). Psychophysical and Theoretical Identifications. *Australasian Journal of Philosophy*, vol. 50, pp 249-258.
- [18] Nagel, E. (1961). *The Structure of Science*, London: Routledge and Kegan Paul.
- [19] Port, R.F., Gelder, T. van (eds.), *Mind as Motion: Explorations in the Dynamics of Cognition*. MIT Press, Cambridge, Mass, 1995.
- [20] Ramsey, F.P. (1929). *Theories*. In: Ramsey, F.P. (1931) *The Foundations of Mathematics and Other Essays* (R. B. Braithwaite, ed.), Routledge and Kegan Paul.
- [21] Rao, A.S. & Georgeff, M.P. (1991). Modelling Rational Agents within a BDI-architecture. In: Allen, J., et al. (eds.), *Proc. of the 2<sup>nd</sup> Intern. Conf. on Principles of Knowledge Representation and Reasoning (KR'91)*, Morgan Kaufmann, pp. 473-484.
- [22] Schoenfield, J.R. (1967). *Mathematical Logic*. Addison-Wesley.
- [23] Steels, L. & Brooks, R. (1995). *The artificial life route to artificial intelligence: Building embodied, situated agents*. Erlbaum.
- [24] Tarski, A., Mostowski, A., and Robinson, R.M. (1953). *Undecidable Theories*. North-Holland.
- [25] Treur, J., Laws and Makeups in Context-Dependent Reduction Relations. In: Love, B.C., McRae, K., and Sloutsky, V.M. (eds.), *Proc. of the 30th Annual Conference of the Cognitive Science Society, CogSci'08*. Cognitive Science Society, Austin, TX, 2008, pp. 1752-1757.