Mental models in the brain: On context-dependent neural correlates of mental models

Jan Treur

Social AI Group, Vrije Universiteit Amsterdam, Netherlands

ARTICLE INFO

Keywords:
Brain
Realisation
Mental model
Neural correlate

ABSTRACT

In this paper, the concept of context-dependent realisation of mental models is introduced and discussed. Literature from neuroscience is discussed showing that different types of mental models can use different types of brain areas. Moreover, it is discussed that the same occurs for the formation and adaptation of mental models and the control of these processes. This makes that it is hard to claim that all mental models use the same brain mechanisms and areas. Instead, the notion of context-dependent realisation is proposed here as a better manner to relate neural correlates to mental models. It is shown in some formal detail how this context-dependent realisation approach can be related to well-known perspectives based on bridge principle realisation and interpretation mapping realisation.

1. Introduction

Mental models can occur in various forms; e.g., (Craik, 1943; Evans, 2006; Furlough & Gillan, 2018; Gentner & Stevens, 1983; Halford, 1993; Johnson-Laird, 1983; Treur & van Ments, 2022). They are a kind of structures or processes in the mind that reflect structures or processes in the world or in other persons. For example, you perceive an impressive course of events in front of you and after closing your eyes you see a kind of movie replay in your mind that replays this course of events. Humans often use some form of mental model to handle situations, for example, operating a device or machine, or to handle somebody else. All such examples show the wide variety of mental models.

A natural question to ask, is about neural correlates of mental models in the brain. How are mental models and their operations encoded as brain states and processes? However, the concept of mental model and the processes in which they are involved have a very diverse appearance in the literature and also the definition and boundaries of the concept mental model are not very sharp. Nevertheless, it is still fair to assume that mental models provide a form of conceptualisation and interpretation of what actually goes on in the brain. But that all diverse types of mental models described in the literature relate in a uniform manner to the same brain states and processes, is probably asked too much. Note that for the sake of simplicity, here the word brain is used while in addition also parts of the body or even in the external world (for example, drawings or notes on paper or on a screen) may be involved in the underlying physically embodied processes.

In this paper, first in Section 2 some literature from neuroscience is discussed where it is shown that different types of mental models can use different types of brain areas, for example in relation to different modalities addressed by a given mental model. Next, in Section 3 from the perspective of philosophy of mind the concept of context-dependent realisation of mental states is discussed. It is illustrated for two well-known cases of multiply realisable concepts: a unified cognitive BDI-model applied to humans and bacteria and the unified notion of force in physics with its different types of realisations. Then, in Section 4 it is discussed how this notion of context-dependent realisation can be applied to mental models. In Section 5 the approach is formalised via two well-known perspectives on realisation from philosophy of mind and philosophy of science: bridge principle realisation and interpretation mapping realisation. Finally, Section 6 is a discussion.

2. Literature on neural correlates for mental models

In this section it is discussed how in the neuroscientific literature various neural correlates for mental models are proposed.

2.1. Some literature from neuroscience

In neuroscience literature, a few examples of how mental models relate to processes in the brain are:
As a first example of the latter, in (Alfred et al., 2020) it is reported that for transitive reasoning, some parts in the brain that relate to spatial representations are also active during activation of abstract mental models concerning abstract objects in the context of an abstract linear order structure (mathematically spoken). Patterns representative of mental models for such examples of linear order structures were revealed in both superior parietal lobule and anterior prefrontal cortex. To get a more general picture, it would be interesting to perform similar experiments for cases where the examples of mental models used do not relate to a linear order structure, as conceptually and mathematically spoken linear order structures are close to the abstract geometric concept of line and therefore these structures and spatial structures are not that far apart.

In (Holyoa & Monti, 2020), considering analogical reasoning, the following is reported indicating that the neural correlates include:

- posterior parietal cortex, implicated in the representation of first-order relations
- rostrolateral PFC, apparently central in integrating first-order relations so as to generate and/or evaluate higher-order relations (e.g., A:B::C:D)
- dorsolateral PFC, involved in maintaining relations in working memory
- ventrolateral PFC, implicated in interference control (e.g., inhibiting salient information that competes with relevant relations)

Here higher-order relations A:B::C:D describe how a first-order relation A:B relates to another first-order relation C:D, as considered in analogical reasoning: A relates to B like C relates to D; for example, ‘dress is to closet as milk carton is to refrigerator’ or ‘shoe is to foot like glove is to hand’; and semantic problems, and that the development of analogical reasoning is associated with increased engagement of the left anterior inferior prefrontal cortex.

2.2. Internal simulation

Another area that addresses brain structures and processes related to mental models is the area of internal simulation. Internal simulation is a very central concept for mental models, especially the ones considering dynamics, as also discussed in (Van Ments & Treur, 2021). It is a means for prediction of the effects of a considered or prepared action without executing it. The idea of internal simulation is that in a certain context (which may cover sensed aspects of the external world, but also internal aspects such as the own goals), preparation states for actions or bodily changes are activated, which, by prediction links, in turn activate certain sensory representation states. The latter states represent the (predicted) effects of the prepared actions or bodily changes, and can be activated from the preparation states by internal connections without actually having executed these actions or bodily changes in the external world or in the body. The notion of internal simulation has been put forward, among others, for:

- prediction of effects of one’s own prepared motor actions; e.g., (Becker & Fuchs, 1985)
- imagination and conscious thought; e.g., (Hesslow, 2002; Hesslow, 2012)
- predicted body states related to preparations for emotional responses, forming a basis for feeling the emotion; e.g., (Damasio 1994, 2003; Bechara & Damasio, 2005)
- recognition or reading another person’s mind, for example, the other person’s emotions; e.g., (Goldman, 2006; Iacoboni, 2008)

As another example, by religious humans a mental God-model is simulated for influencing their behaviour as also addressed in (Van Ments, Treur, & Roelofsema, 2018; Van Ments, Treur, & Roelofsema, 2022). This mental God-model refers to the personal God of the individual. As discussed in (Kapogiannis et al., 2009; Kapogiannis, Deshpande, Krueger, Thornburg, & Grafman, 2014; Schaap-Jonker, Sizoo, van Schotter-van Roekel, & Corveleyn, 2013), this mental God-model consists of both an emotional part and a cognitive part, and both parts are dynamically interrelated. The emotional part is unconsiously developed, highly influenced by parents and significant others. The emotional and the cognitive part that form the mental God-model can be related to different parts in the brain as studied, for example, by the above-mentioned (Kapogiannis et al., 2009; Kapogiannis et al., 2014; Schaap-Jonker et al., 2013). The emotional part involves

- the amygdala, basal ganglia,
- the ventromedial prefrontal cortex, the lateral temporal cortex,
- the dorsal anterior cingulated cortex, and the orbitofrontal cortex.

These parts of the brain are involved in assigning emotional significance to behaviour and events and to control of cognition and emotion. Moreover, the cognitive part of the mental God-model involves

- the lateral prefrontal cortex, the medial prefrontal cortex,
- the lateral parietal cortex, the medial parietal cortex,
- and the medial temporal lobe

These all are brain circuits that more in general are responsible for the processing of more complex linguistic and symbolic input. For the case of mental God-models considered here, the above indicated combination of brain parts enable the formation of the personal mental God-model of the individual.

All such types of internal simulation use internal connections or causal pathways from an action preparation state to some type of sensory representation state for the (predicted) effect of this action (without actually executing the action). Such relations and processes are often part of mental models. For example, Damasio calls such pathways (in particular, to generate feelings) as-if body loops (Damasio, 1994; Damasio, 2003; Bechara & Damasio, 2005), while Hesslow (2002) refers to them (considering a more general context) as ‘simulation of behaviour and perception’ or simulated perception-behaviour chains. For both types of causal pathways, see Fig. 1. In the latter case the emphasis is on longer chains, as every sensory action effect representation can trigger preparation for a new action, which in turn can trigger a new predicted sensory action effect representation, and so on. These chains are proposed by Hesslow (2002) as the neural basis for conscious thought.

Such structures of pathways for internal simulation are realisations in the mind of mental models that are executed. In case these mental models relate to processes in someone else’s mind, these chains refer to the mind of the other person, like in ‘simulating minds’ by which mindreading can be achieved in combination with mirroring (Goldman, 2006; Iacoboni, 2008) or in Theory of Mind. In Fig. 1 two original pictures of as-if body loops (Damasio, 1994; Damasio, 2003; Bechara & Damasio, 2005) and of simulated perception-behaviour chains (Hesslow, 2012) illustrate the idea of internal simulation in some more detail.
Viewed from a higher abstraction level, all these different types of processes in the brain serve as some form of internal simulation. However, in these different cases, different brain states, pathways and areas are used. For example, mental models involving emotions and feeling states associated to some considered action or belief (i.e., mental models involving an emotional context), will use parts and pathways of the brain that are not the same as mental models that do not involve such emotions and feeling states (i.e., mental models involving a non-emotional context).

Right picture, adopted from (Hesslow, 2012): (a) Stimulus $S_1$ causes perceptual activity $s_1$, which causes preparatory response $r_1$ and overt response $R_1$. This $r_1$ causes predictable new stimulus $S_2$, which causes new sensory activity, etc. (b) Preparatory response $r_1$ elicits, via internal association mechanisms, perceptual activity $s_2$ before overt behaviour occurs and causes new stimulus.

The notion of internal simulation can be viewed as an abstraction that unifies these different types of brain processes. More in general, the neural circuits to internally simulate processes from the external world will be different from the circuits used when simulating mental processes of other persons. Such simulations will usually apply the same brain structures as those involved in perceiving the processes in reality; for example, perceiving the own or someone else’s body states uses brain areas that are different from brain areas used when perceiving states of the physical world.

2.3. Neural correlates for adaptation and control for mental models

From the above it seems that most research on the neuroscience of mental models focuses on the use of mental models and not on their formation, adaptation or control as discussed, for example, in (Van Ments & Treur, 2021). For the latter types of processes, still other parts and pathways in the brain may be used. For formation and adaptation of mental models, the extensive neuroscience literature on plasticity may be relevant, such as (Hebb, 1949; Chandra & Barkai, 2018; Daoudal & Debanne, 2003; Debanne, Inglebert, & Rizzoli, 2019; Sjöström, Rancz, Roth, & Hauser, 2008) to name just a few. For control, probably some parts of the prefrontal cortex concerning executive functions and cognitive control may be involved, but also literature on the more detailed neuroscience of metaplasticity for control of plasticity such as (Abraham & Bear, 1996; Magerl, Hansen, Treede, & Klein, 2010) may be relevant. So, there are still some challenges left to be explored for the area of neural correlates for mental model handling.

3. Context-dependent realisation of mental states

As discussed above, proposed neural correlates for mental models show a diversity of occurrences. This does not fit well to a maybe preferred option that there is one universal mechanism in the brain that realises all mental models. Perhaps it is asked too much to assume that there is one fixed architecture in the brain that realises all types of mental models. This suggests that other options may be considered that fit better. Within philosophy of mind, from a wider context a similar issue is addressed: the issue of multiple realisability of mental states; e.g., (Kim, 1996). Here an interesting option to address this issue is discussed, namely the perspective based on context-dependent realisation. This looks like a more promising perspective than assuming that one universal brain structure can be found as a correlate for handling all types of mental models.

3.1. Context-dependent multiple realisation of mental states

According to this alternative perspective, instead of a one-to-one correspondence of all types of mental models to one specific type of brain structure, a more realistic approach is by relating mental models to brain areas in a more pluriform and context-related manner. In particular, the notion of context-dependent multiple realisation as suggested by Kim (1996), pp. 233–236, can provide a useful way of interpretation of the situation. Here, roughly spoken, depending on the context a mental state can relate to different types of brain states and processes (multiple context-specific realisations can exist), and within each context the specific causal relations for these brain states should be in accordance with the relations assumed for the considered mental states. A context is here, for example, the physical makeup of an organism. These makeups usually differ for different species and individuals, but at a more abstract level still the same mental concepts can be used to describe them in a unified manner. More details about this perspective of context-dependent realisation (and how this can be used more generally to clarify how mental relations or laws and neurological relations or laws relate to each other) can be found in (Treur, 2008; Treur, 2011).

Based on context-dependent realisation, the mental states and their assumed causal relations form a unified high-level description of a number of different specific brain states and their specific causal relations. For example, suppose mental states $M$ and $M'$ are considered with an assumed causal relation $M \rightarrow M'$; see Fig. 2. Then, for example, in two different contexts $C_1$ and $C_2$ two different types of realisations may be considered, one in context $C_1$ where $M$ is realised by brain state $B_1$ and $M'$ by brain state $B'_1$ and another one in context $C_2$ where $M$ is realised by brain state $B_2$ and $M'$ by brain state $B'_2$. Then, for a faithful realisation it is required that causal relations $B_1 \rightarrow B'_1$ within context $C_1$ and $B_2 \rightarrow B'_2$ within context $C_2$ exist between these brain states. In this case, at a higher, more abstract level of description the causal relation $M \rightarrow M'$ unifies these specific causal relations $B_1 \rightarrow B'_1$ and $B_2 \rightarrow B'_2$ within the two different contexts, as shown in Fig. 2. In Sections 3.2 and 3.3 some examples of multiple realisation are presented; in Section 5 a formalisation is addressed.
3.2. An illustration from biology: Multiple realisation of behavioural choice

One illustration, borrowed from the work described in (Jonker, Snoep, Treur, Westerhoff, & Wijngaards, 2002; Jonker, Snoep, Treur, Westerhoff, & Wijngaards, 2008) is the following (see Fig. 3). Here the left-hand side describes a causal network for how an E. coli bacterium determines what food it uses as intake (according to the literature in biochemistry) and the right-hand side describes a causal network for how a human is assumed to do that (according to the so-called BDI-model). The horizontal dashed double arrows show how the states for DNA, mRNA, active enzyme and flux of an E. coli correspond to states for desire, intention, readiness, and action, respectively for a human.

Similar correspondences can be made for the other nodes in the two networks as indicated by the longer dashed double arrows. This example shows how the BDI-model (originally meant for human mental processes and behaviour) can also be used as a more general unified description of mental processes, unifying processes in different types of organisms with different physical makeups where the general unified model gets its different context-dependent realisations.

The perspective discussed above is just one example of a form of unification: different types of processes are comparable, and we can, for example, compare the processes underlying human intelligence and behaviour to the processes underlying bacterial behaviour, as described from a wider perspective in (Jonker et al., 2002; Jonker et al., 2008; Westerhoff, He, Murabito, Créamy, & Barberis, 2014a; Westerhoff, Brooks, Simeonidis, García-Contreras, He, Boogerd, Jackson, Gongcharuk, & Kolodkin, 2014b). For example:

“We have become accustomed to associating brain activity – particularly activity of the human brain – with a phenomenon we call “intelligence.” Yet, four billion years of evolution could have selected networks with topologies and dynamics that confer traits analogous to this intelligence, even though they were outside the intercellular networks of the brain. Here, we explore how macromolecular networks in microbes confer intelligent characteristics, such as memory, anticipation, adaptation and reflection and we review current understanding of how network organization reflects the type of intelligence required for the environments in which they were selected. We propose that, if we were to leave terms such as “human” and “brain” out of the defining features of “intelligence,” all forms of life – from microbes to humans – exhibit some or all characteristics consistent with “intelligence”. (Westerhoff et al., 2014b), p. 1.

This quote emphasizes that not only in the human brain, but even in the smallest life forms many if not all aspects of intelligence as usually attributed to humans are realised in a variety of different manners using different types of mechanisms and causal relations underlying them.

3.3. An illustration from physics: Multiple realisation of force

Context-dependent multiple realisation can also be found in other domains, for example, for the notion of force within physics, as described by Nagel (1961, pp. 186-192); see also (Treur, 2007). Force is a general concept that unifies multiple occurrences of specific forces in different contexts. Depending on the context defined by a considered world configuration, one type of realisation of a force is by gravitation, but other types are forces realised by electrical charges, by magnetic objects, or by deformation caused by collisions, or gas temperature, for example. All these different types of realised forces (1) are generated through different mechanisms based on different types of causal relations (Nagel calls these ‘force functions’), but (2) in a unified manner have exactly the same effect on the acceleration a of an object with mass m according to the well-known law $F = ma$ which relates force F to acceleration a. The successfulness of this law illustrates within this physical domain the power of the idea of a unified concept with multiple realisations.

4. Context-dependent realisation of mental models

Now, returning to mental models, suppose as part of a mental model a relation $M \rightarrow M'$ is assumed. If the idea of context-dependent realisation discussed in Section 2 is applied to mental models, then similar to the above mental concepts $M$ and $M'$ and their causal relation, this idea can be applied to any mental model relation $M \rightarrow M'$; then the left hand picture shown in Fig. 4 is obtained for such a mental model relation. Here contexts such as $C_1$ and $C_2$ may depend on the type of species or person and the type of mental model that is considered. This means that as within the given mental model, $M$ and $M'$ relate according to $M \rightarrow M'$, and as corresponds to $B_1$ and $M'$ to $B'_1$ within context $C_1$, for a faithful realisation there should be a relation $B_1 \rightarrow B'_1$ within that context, and similarly a relation $B_2 \rightarrow B'_2$ for context $C_2$ and $B_2$ and $B'_2$. Note that here it is assumed that the relations within a mental model can be of any type of relation, causal or not. Then they have to correspond accordingly to certain kinds of relations in the brain. If in the mental model the relations considered are meant as causal relations, then the corresponding relations in the brain can also be taken as causal.

Fig. 3. Multiple realisations of a general unified BDI-model for mental processes in an E. coli bacterium (left hand side) and in a human (right hand side) and their mutual correspondence relations (horizontal dashed double arrows).
relations. This causality can still be of many different forms, for example, varying from a description of successive relational or analogical reasoning steps to algorithmic steps in an algorithmic skill or any (other) type of causality underlying dynamical systems.

As a mental model can also be assumed to relate to actual relations in the world (e.g., see Van Ments & Treur, 2021), at the same time the right-hand picture in Fig. 4 applies, where relation $W \rightarrow W'$ in the world corresponds to relation $M \rightarrow M'$ in the mental model. Then the assumption that the mental model relations $M \rightarrow M'$ correspond to relations $W \rightarrow W'$ in the world plus the assumption that mental model relations $M \rightarrow M'$ correspond to (for example) relations $B_2 \rightarrow B'_2$ between states in the brain within context $C_2$ imply by transitivity of ‘correspondence’ that these relations $B_2 \rightarrow B'_2$ in the brain also correspond to the relations $W \rightarrow W'$ in the world (see the dashed arrows between the left and right picture in Fig. 4). That means that the brain processes simulate the world processes according to similar relations, which is in line with (Craik, 1943). In his book (Craik, 1943) he describes a mental model as a small-scale model that is carried by an organism within its head and used to try out alternatives of actions before executing them as follows:

‘... it is a physical working model which works in the same way as the process it parallels...Thus, the model need not resemble the real object pictorially; Kelvins’ tide-predictor, which consists of a number of pulleys on levers, does not resemble a tide in appearance, but it works in the same way in certain essential respects...’ (Craik, 1943, p. 51).

‘If the organism carries a “small-scale model” of external reality and of its own possible actions within its head, it is able to try out various alternatives, conclude which is the best of them, react to future situations before they arise, utilise the knowledge of past events in dealing with the present and future, and in every way to react in a much fuller, safer, and more competent manner to the emergencies which face it.’ (Craik, 1943, p. 61)

He emphasizes that such internal models work in a way similar to how the real world works.

In Fig. 4, for the sake of simplicity and explanation only one mental model relation is considered. As in general a mental model involves a whole network of such relations, a more realistic picture is shown in Fig. 5.

Here for a faithful realisation, all relations in the mental model network have to correspond to similar relations in the brain and for a
faithful representation of the world all relations in the mental model network have to correspond to similar relations in the world. As a result, the corresponding network in the brain will faithfully simulate the world processes.

Note that the perspective based on context-specific realisation allows to maintain a very general notion of mental model unifying all types of mental models, also those that use very different brain processes. But within that general notion of mental model, as a form of classification still specific types of mental models can be considered, for example types of mental models that do share a common structure for their realisation in the brain. In a sense, this provides the best of two worlds: (1) there is one universal notion of mental model with general knowledge and theory covering a very wide variety of cases, and (2) under the umbrella of this general notion of mental model, still a number of very specific types of mental models can be studied as well with more specific knowledge and theories in addition. In Section 3 a few specific results on neural correlates of different types of mental models will be discussed that might be considered to provide some evidence in favour of this perspective of context-specific realisations.

5. Context-dependent realisation from different perspectives

Based on the notion context-dependent realisation as introduced in (Treur, 2008), a set of contexts can be identified and realisations of mental models can be related to these contexts. Assuming that contexts are defined in a sufficiently fine-grained manner, within one context the realisation is unique. Then contexts can be seen as a form of parameterisation of the realisations. For mental models, for example, these contexts may be based on different types of sensory representations. For a context-dependent realisation approach, a (neurological) background base theory T is assumed with a set of contexts C such that each particular context is formally described by a context S ∈ C. The contexts S are assumed to be descriptions in the language of T and consistent with T. The contexts S ∈ C can be used to distinguish the different realisations that are possible for mental models. This means that for a given mental model a context S can be found such that all relations of the mental model can be related to realisers within this context S. Below it is shown how this context-dependency can be addressed for two well-known general approaches to realisation, namely bridge principle realisation (Nagel, 1961) and realisation by an interpretation mapping; e.g., (Bickle, 1992) and ( Hodges, 1993), pp. 201–263. Here a fixed (neurological) background theory T is assumed. It will be assumed as a general setting that a mental model is defined by a set of relations R(a1, ..., ak) between basic concepts ai. For example, in Fig. 4, such a relation R is denoted by an arrow; for the example mental model depicted in Fig. 5, in the R-notation the relations are R(M1, M2), R(M3, M5), R(M2, M3), and so on.

5.1. Context-dependent bridge principle realisation

For the bridge principle realisation approach, for a given relation R(a1, ..., ak) the set of realisers that exists within a context S ∈ C, is expressed by context-dependent biconditional bridge principles parameterised by context S ∈ C, specified by

\[ a_1 \leftrightarrow b_{1 S}, \ldots, a_k \leftrightarrow b_{k S} \]

In Fig. 5, these correspond to the blue dashed double arrows, so they can be specified by:

\[ M_1 \leftrightarrow B_1, M_2 \leftrightarrow B_3, M_3 \leftrightarrow B_1 \]
\[ M_4 \leftrightarrow B_1, M_1 \leftrightarrow B_3, M_5 \leftrightarrow B_5 \]

Given such a specification, context-dependent bridge principle realisation within context S for the relations R(a1, ..., ak) defining a given mental model can be formulated in two equivalent manners by (where \( \vdash \) is a symbol for logical entailment):

(i) \[ R(a_1, \ldots, a_k) \vdash T \cup S \cup (a_1 \leftrightarrow b_{1 S}, \ldots, a_k \leftrightarrow b_{k S}) \Rightarrow \]

(ii) \[ R(a_1, \ldots, a_k) \vdash T \cup S \Rightarrow (b_{1 S}, \ldots, b_{k S}) \]

Note that context-dependent bridge principle realisation implies unique realisers (up to equivalence) per context S: from \( a_1 \leftrightarrow b_5 \) and \( a \leftrightarrow b_3 \) it follows that \( b_2 \) and \( b_5 \) cannot be non-equivalent in S. So to obtain context-dependent bridge principle realisation in cases of multiple realisation, the contexts are defined with a grain-size such that per context a unique realisation exists.

5.2. Context-dependent interpretation mapping realisation

A context-dependent interpretation mapping is a multi-mapping of concepts parameterised by contexts: a multi-mapping \( \varphi_S (S \in C) \) from mental model concepts to concepts of the background (neurological) theory parameterised by contexts S ∈ C. For example, in Fig. 5, following the blue dashed double arrows, such a mapping can be defined by:

\[ \varphi_1 (M_1) = B_1, \varphi_1 (M_2) = B_2, \varphi_1 (M_3) = B_3 \]
\[ \varphi_2 (M_4) = B_4, \varphi_2 (M_5) = B_5, \varphi_2 (M_6) = B_6 \]

These mappings are assumed compositional in the sense that for any mental model relation \( R(a_1, \ldots, a_k) \) it is assumed

\[ \varphi_b (R (a_1, \ldots, a_k)) = R (\varphi_1 (a_1), \ldots, \varphi_1 (a_k)) \]

Such a multi-mapping is a context-dependent interpretation mapping realisation when it satisfies the property that for any context S ∈ C for any relation \( R(a_1, \ldots, a_k) \) in a given mental model, the relation \( \varphi_S (R(a_1, \ldots, a_k)) \) is entailed by S:

\[ R(a_1, \ldots, a_k) \vdash T \cup S \Rightarrow \varphi_S (R(a_1, \ldots, a_k)) \]

5.3. Relating bridge principle realisation and interpretation mapping realisation

In this section it is shown how context-dependent bridge principle realisation can be translated into context-dependent realisation based on an interpretation mapping and vice versa.

5.3.1. From interpretation mapping realisation to bridge principle realisation

Suppose a context-dependent interpretation mapping realisation \( \varphi_S \) is given for some \( S \in C \). For each basic concept \( a_i \) of a mental model, specify the bridge principle

\[ a_1 \leftrightarrow b_{1 S}, \ldots, a_k \leftrightarrow b_{k S} \]

If \( R(a_1, \ldots, a_k) \) is mental model relation involving concepts \( a_1, \ldots, a_k \), then

\[ T \cup S \Rightarrow \varphi_S (R(a_1, \ldots, a_k)) \]

By compositionality of mapping \( \varphi_S \) it follows that

\[ T \cup S \Rightarrow R(\varphi_S (a_1), \ldots, \varphi_S (a_k)) \]

Therefore it follows

\[ T \cup S \Rightarrow R(b_{1 S}, \ldots, b_{k S}) \]

This shows that the criterion for context-dependent bridge principle realisation within context S is fulfilled.

5.3.2. From bridge principle realisation to interpretation mapping realisation

For a translation the other way around, assume for context-dependent bridge principle realisation, for some \( S \in C \) bridge principles

\[ a_1 \leftrightarrow b_{1 S}, \ldots, a_k \leftrightarrow b_{k S} \]

are given for the basic concepts \( a_i \) of a mental model such that the bridge principle realisation criterion for context S and bridge principles
Define the mapping $\phi_S$ for each basic expression $a_i$ based on the given bridge principle $a_i \leftrightarrow b_{S,i}$, by

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

For $R(a_1, \ldots, a_k)$ extend this by compositionality

$$\phi_S(R(a_1, \ldots, a_k)) = R(\phi_S(a_1), \ldots, \phi_S(a_k))$$

For this mapping $\phi_S$, from $R(a_1, \ldots, a_k)$ by the bridge principle realisation criterion it follows:

$$R(a_1, \ldots, a_k) \Rightarrow T \cup S-R(b_1, \ldots, b_S)$$

Therefore, the criterion for a context-dependent interpretation mapping realisation is fulfilled. Note that the translations from context-dependent bridge principle realisation to context-dependent interpretation mapping realisation and from context-dependent interpretation mapping realisation to context-dependent bridge principle realisation as given are each other’s inverse.

6. Discussion

In this paper, the use of the concept of context-dependent realisation of mental states from philosophy of mind for mental models was discussed. This concept was illustrated for two wellknown cases of multiply realisable concepts: a unified cognitive BDI-model applied to humans and bacteria and the unified notion of force in physics with its different types of realisations. As the core of this paper, it was discussed how this idea of context-dependent realisation can be applied to mental models.

Some literature from neuroscience was discussed where it is shown that different types of mental models can use different types of brain areas. For example, some types of mental models address spatial or linearly ordered structures and turn out to make use of brain areas that typically relate to the processing of spatial information; e.g., (Alfred et al., 2020). Other examples of mental models may concern emotions of other persons; these mental models turn out to make use of brain parts typically involved in emotions and feelings; e.g., (Damasio, 1994; Iacoboni, 2008). Moreover, it was discussed that this diversity also applies to the formation and adaptation of mental models and the control of these processes. This makes that the notion of context-dependent realisation can be a suitable manner to relate neural correlates to mental models in a pluriform manner. This has been worked out more formally in Section 5.

More specifically, these observations suggest a perspective on context-dependent neural correlates of mental models where this context-dependency actually concerns the type of content of the mental model: what it represents. It might be regretted that in this way these neural correlates do not concern one universal mechanism in the brain that handles all mental models. However, in the paper it has been shown that the notion of context-dependent realisation from Philosophy of Mind (Kim, 1996) still provides a neat foundational description of this more pluriform perspective. In addition, it has been discussed that also in other scientific disciplines this perspective occurs; for example, not only for mental states in general as put forward by Kim (1996), but within physics the notion of force $F$ used in the very successful law $F = ma$ relating force to acceleration $a$, also has multiple context-dependent realisations by essentially different (physical) mechanisms such as gravitation, electrical charge, magnetic influence, deformation by collision, gas temperature, etc. (Nagel, 1961). Therefore, the topic of mental models is in good scientific company concerning this perspective of context-dependent realisation.

Finally, this idea has some relation to the historical Simulation-Theory versus Theory-Theory discussion for understanding each other’s minds; e.g., (Goldman, 2006), pp. 10–22. From a Theory-Theory perspective it may be tempting to look for one universal (‘reasoning’) mechanism in the brain to reason with all theories (mental models) of others’ minds. But from a Simulation-Theory perspective, it makes more sense that brain areas for various types of modalities are used for internal simulation of theories (mental models) of others’ minds, and these modalities correspond to the modalities for the content of the mental model at hand. In that sense the proposed perspective on context-dependent neural correlates for mental models may relate more to the Simulation-Theory perspective than to the Theory-Theory perspective.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abraham, W. C., & Bear, M. F. (1996). Metaplasticity: the plasticity of synaptic plasticity. Trends in Neuroscience, 19(4), 126–130.
Alfred, K. L., Connolly, A. C., Cetron, J. S., et al. (2020). Mental models use common neural spatial structure for spatial and abstract content. Communications Biology, 3, 1–17.
Becker, W., & Fuchs, A. F. (1985). Prediction in the Oculomotor System: Smooth Pursuit During Transient Disappearance of a Visual Target. Experimental Brain Research, 57 (1985), 562–575.
Bechara, A., & Damasio, A. R. (2005). The somatic marker hypothesis: A neural theory of economic decision. Games and Economic Behavior, 52(2005), 336–372.
Bickle, J. (1992). Mental Anatomy and the New Mind-Brain Reductionism. Philosophy of Science, 59, 217–230.
Chandra, N., & Barkai, E. (2018). A non-synaptic mechanism of complex learning: Modulation of intrinsic neuronal excitability. Neurobiology of Learning and Memory, 154, 30–36.
Cohen, A. J., Levitin, D., & Kleber, B. (2020). Brain mechanisms underlying singing. In P. A. Russo, B. Bari, & A. J. Cohen (Eds.), Routledge Companion to Interdisciplinary Studies in Singing: Vol I, Development (pp. 79–86). Routledge.
Craik, K. J. W. (1943). The nature of explanation. Cambridge, MA: University Press.
Debanne, D., & Debanne, D. (2003). Long-term plasticity of intrinsic excitability: Learning rules and mechanisms. Learning, Mem., 10, 456–465.
Debanne, D., Inglebert, Y., & Russier, M. (2019). Plasticity of intrinsic neuronal excitability. Current Opinion in Neurobiology, 54, 73–82.
Evens, J. (2006). The heuristic-analytic theory of reasoning: Extension and evaluation. Psychonomic Bulletin & Review, 13(3), 378–395.
Furlough, C. S., & Gillan, D. J. (2018). Mental Models: Structural Differences and the Role of Experience. Journal of Cognitive Engineering and Decision Making, 12(4), 269–287. https://doi.org/10.1177/1553608417733256.
Garvert, M., Dolan, R., & Behrens, T. (2017). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. eLife, 6, pii: e17086.
Gentner, D., & Stevens, A. L. (2002). Mental models. Hilldale NJ: Erlbaum (1983).
Goldman, A. I. (2006). Simulating Minds: The Philosophy, Psychology, and Neuroscience of Mindreading (p. 2006). New York: Oxford University Press.
Halford, G. S. (1993). Children’s understanding: The development of mental models. Lawrence Erlbaum Inc.
Hebb, D. O. (1949). The organization of behavior: A neuropsychological theory. Wiley.
Heslow, G. (2002). Conscious thought as simulation of behaviour and perception. Trends in Cognitive Sciences, 6(2002), 242–247.
Heslow, G. (2012). The current status of the simulation theory of cognition. Brain Research, 1428(2012), 71–79.
Hodges, W. (1993). Model theory. Cambridge University Press.
Horyouk, K. J., & Monti, M. M. (2020). Relational Integration in the Human Brain: A Review and Synthesis. Journal of Cognitive Neuroscience.
Hurley, S. (2008). The shared circuits model (SCM): How control, mirroring, and simulation can enable imitation, deliberation, and mindreading. Behavioral and Brain Sciences, 31(1), 1–22.
Iacoboni, M. (2008). Mirroring People: The New Science of How We Connect with Others. New York: Farrar, Strauss and Giroux.
Johnson-Laird, P. N. (1983). Mind: Science, Models, and Simulations. Cambridge University Press.
Jonker, C. M., Snoep, J. L., Treur, J., Westerhoff, H. V., & Wijngaards, W. C. A. (2002). BDI-modelling of complex intracellular dynamics. J. of Theoretical Biology, 214, 1–23.
Kapogiannis, D., Barbey, A. K., Su, M., Zamboni, G., Krueger, F., & Grafman, J. (2009). Cognitive and neural foundations of religious belief. Proceedings of the National Academy of Sciences, 106(12), 4876–4881.


