A Neural Dynamic Architecture That Autonomously Builds Mental Models

Parthena Kounatidou (parthena.kounatidou@ini.rub.de)
Mathis Richter (mathis.richter@ini.rub.de)
Gregor Schöner (gregor.schoener@ini.rub.de)

Institut für Neuroinformatik, Ruhr-Universität Bochum, 44780 Bochum, Germany

Abstract

Reasoning and other mental operations are believed to rely on mental models. Arguments have been made that mental models share representational substrate with perception. Here, we demonstrate that a neural dynamic architecture that perceptually grounds language may also support the building of mental models. Supplied with a sequence of simple premises that specify the colors of object pairs as well as their spatial relation, the architecture builds a mental model of the described scene. We show how the neural processes of the architecture evolve in response to both determinate and indeterminate premises. For indeterminate premises, we demonstrate that the preferred mental models observed in human participants emerge from the underlying neural dynamics.

Keywords: mental models; neural dynamics; dynamic field theory; grounded cognition; visual imagery

Introduction

Most of our thinking relates to real or imagined scenes and events. Mental models of such scenes can be built from language and enable us to understand statements, reason about them, and make inferences (Johnson-Laird, 2010; Knauff, 2013). But how do neural processes generate such mental models?

Mental models may have propositional form, but may also entail perceptual representations (Mani & Johnson-Laird, 1982; Zwaan, 2014) and, therefore, mental imagery. We know that when we imagine something, the brain's sensorimotor areas are engaged as if we were actively perceiving a scene (Kosslyn, Ganis, & Thompson, 2001; Pulvermüller, 2005). Many thus believe that mental imagery is based on the same kinds of perceptual representations that underlie perception and cognition in general (Barsalou, 2009; Gallese & Lakoff, 2005). If mental imagery and mental models share their representational format with perception and cognition, then this suggests a route for providing a neural process account. Here we pursue this route by showing that a neural architecture of the perceptual grounding of language (Richter, Lins, Schneegans, Sandamirskaya, & Schöner, 2014; Richter, Lins, & Schöner, 2017) can serve to build mental models.

The neural architecture interfaces continuous perceptual representations with representations of discrete concepts. It is based on dynamic field theory (DFT), a mathematical and conceptual framework for modeling cognitive processes that is consistent with neural principles (Schöner, Spencer, & the DFT Research Group, 2015). In DFT, neural activation is captured by dynamic neural fields that are defined over continuous feature dimensions and evolve in continuous time based on differential equations. This enables the coupling of neural representations to online sensory and motor processes.

The current paper demonstrates how the same architecture, with minimal changes, captures the process of building mental models. The architecture is supplied with multiple premises about objects, specifying their colors and spatial relations. It is able to build a mental model of the described scene based on continuous perceptual representations. This first requires that the discrete color and relational concepts invoked by the description activate continuous perceptual representations, a mapping that we assume is previously learned. Furthermore, it requires that new objects are placed in an imagined scene representation or mental canvas, based on the specified spatial relation to other objects. In our language, translating relations into spatial positions entails active coordinate transforms. Building the scene representation also requires that spatial positions are bound to continuous feature representations, such as color. Ultimately, this bound representation of features and space must yield a stable working memory of the imagined scene. When a sentence refers to an object that is already part of the mental model, it must be brought into the attentional foreground, guided by its distinct features. Finally, the time-continuous neural dynamics of this architecture must organize all of these processes, which entails the generation of sequences.

A particular challenge appears in ambiguous descriptions. Take the following example: "Imagine a green ball. Now imagine a red ball to the left of that green ball. Now imagine a blue ball to the left of the green ball." The first and second premise are *determinate* problems, where the premise unambiguously specifies where to place the new object. The last premise, on the other hand, is an *indeterminate* problem because it does not specify whether the blue object should be placed to the left or to the right of the red object. Ragni and Knauff (2013) find that most people solve indeterminate descriptions by sequentially placing objects such that they create the least amount of change relative to the already established scene. They capture this *preferred mental model* by an architecture that places objects in the first free slot on a grid-canvas. This paper demonstrates how the same behavior may emerge from the neural architecture. The positioning of new objects emerges from inhibitory influences of objects that are already present in the scene.

Methods

Dynamic field theory (DFT) is a mathematical framework for modeling cognitive processes based on neural principles (Schöner et al., 2015). In DFT, the activation of populations of neurons is captured by dynamic neural fields. Fields

are defined over continuous feature dimensions, for example color or space, and evolve in continuous time based on the following integro-differential equation

$$\tau \dot{u}(x,t) = -u(x,t) + h + s(x,t) + \int g(u(x',t))\, w(x-x')\, dx'.$$

Here, $u(x,t)$ is the activation of a field defined over the continuous feature dimension x at time t. τ is a time constant that determines the time scale of the dynamics, h is a negative resting level, and $s(x,t)$ is external input from other fields or sensors. The last term, the integral, formalizes lateral interaction within the field. The interaction kernel w features local excitation and mid-range or global inhibition. The strength of interaction is determined by the field's sigmoidal output function $g(u(x,t))$ with a threshold at zero (Amari, 1977).

With sufficient external input $s(x,t)$, a field goes through a dynamic detection instability, where the subthreshold attractor becomes unstable and a new attractor emerges above threshold. This leads to the formation of a stable peak of activation—the unit of representation in DFT. The shape of the interaction kernel determines whether a field may form multiple peaks or make a selection decision to form only a single peak. With sufficient self-excitation, peaks become self-sustained and remain stable after initial input is removed. Fields may be defined over multiple continuous feature dimensions. Dynamic neural nodes are not defined over any feature dimension and instead represent discrete concepts. Fields and nodes may be coupled to form larger architectures, where fields of different dimensionality are connected via shared feature dimensions. DFT architectures are continuously updated and may be coupled to sensory input and motor output.

Architecture

The architecture introduced here (Figure 1) is composed of dynamic neural fields and nodes that are interconnected to form a single dynamical system. It can be functionally divided into five parts, described in the following sections.

Concepts

The user interacts with the architecture by supplying premises such as "an orange object to the left of a blue object." Each discrete concept of color and spatial relation contained within such a premise is represented by a pair of dynamic neural nodes. *Memory nodes* (blue circles in Figure 1) represent part of the premise and act as an interface for the user. *Production nodes* (pink circles) gate the influence of memory nodes onto the architecture. The perceptual meaning of a concept is encoded in patterned synaptic connections (marked with a star) between the production nodes and a dynamic neural field. Color concepts (i.e., RED, BLUE, CYAN, GREEN, ORANGE) are encoded by connections from their production node to every location of the color attention field, which is defined over the hue dimension. The weights that make up the connection patterns are determined by Gaussians centered on the respective colors. Spatial relational concepts (i.e., TO THE LEFT OF, TO THE RIGHT OF, ABOVE, BELOW) are encoded by connections from their production node to every location of the relational field, which is defined over two-dimensional space. These connection patterns are inspired by empirical data (Logan & Sadler, 1996); they are depicted in Figure 1 next to their respective concept nodes using a color-code. In a relational premise like "the orange object to the left of the blue object", color concepts appear in the roles of the target object (here, orange) and the reference object (blue). In the architecture, each color concept thus appears twice, for target and reference.

Attention

The attentional system consists of two fields. The color attention field is defined over the circular hue dimension. A peak in this field brings objects of the specified color into the attentional focus. It feeds excitatorily into the three-dimensional attention field along their shared hue dimension. The attention field is defined over two additional spatial dimensions that span a canvas on which objects may be imagined. When the color attention field has a peak, its output forms a subthreshold pattern of activation in the attention field in the shape of a sheet. When this sheet of activation overlaps with subthreshold localized input along the spatial dimensions, the field may form a peak. This is visible in Figure 1 in the lowest slice through the attention field. The subthreshold localized input along the spatial dimensions can come from four sources. For objects that are already part of a mental model, the input comes from the scene representation field, which holds a representation of the mental model. For new objects that are to be added to a mental model, the input comes from the target field. In case the object is the first one to be placed in the model, a localized bias input places it at the center of the spatial canvas. This case is detected by the color mismatch field, which forms a peak if the color represented in the color attention field does not already exist in the mental model. A last input to the attention field comes from the reference field and supports inferences about reference objects.

Scene representation

The mental model is built and memorized in the scene representation field, also defined over hue and two-dimensional space. Strong self-excitation as well as surround inhibition ensure that peaks in this field remain stable even after excitatory input from the attention field is removed. The output of the scene representation field feeds excitatorily into the spatial scene representation field. It holds a representation of the spatial positions of all objects in the mental model.

Spatial transformation and object creation

The spatial transformation system enables the architecture to express spatial relational premises. A premise such as "the orange object to the left of the blue object" (shown in Figure 1) consists of three elements, all of which need to be represented by the architecture: the object the premise is primarily referring to (the target object, here orange), the spatial

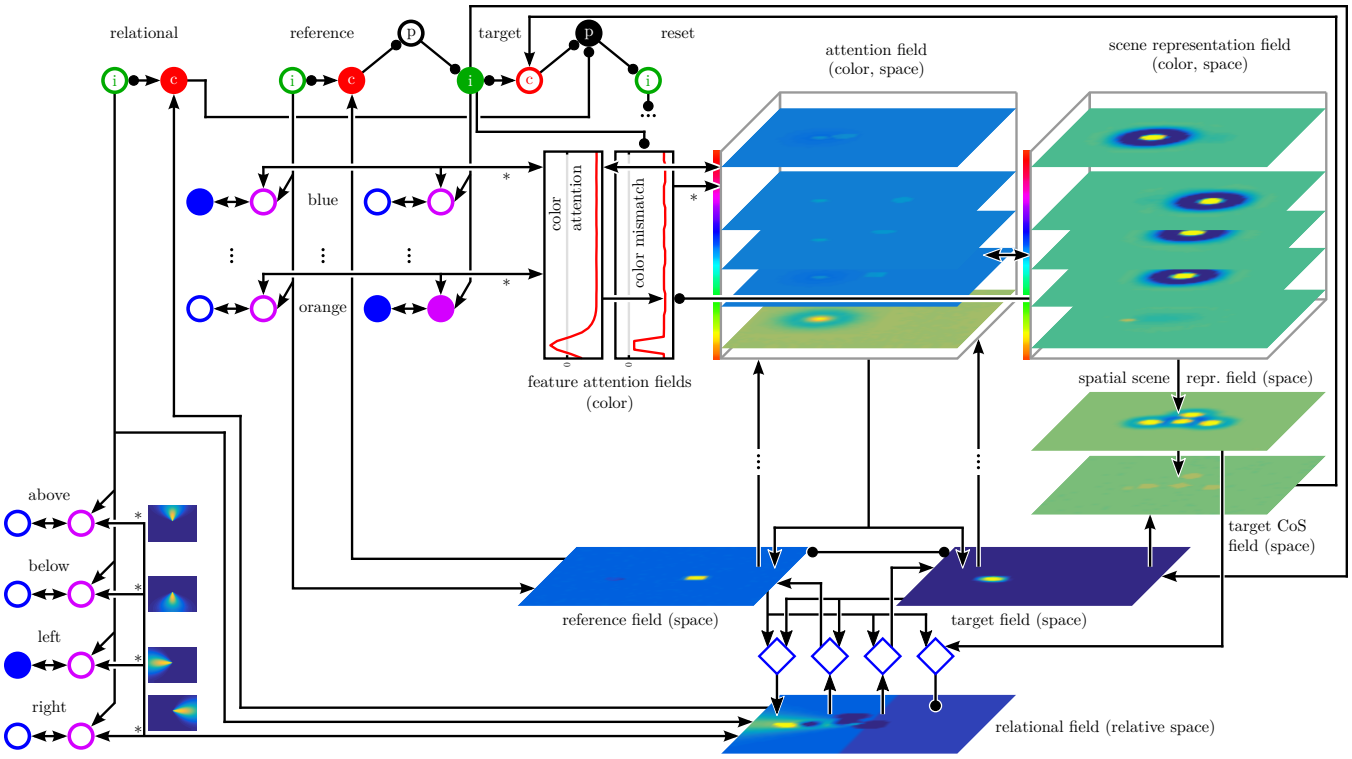

Figure 1: Activation snapshot of the architecture as it forms a mental model consisting of five objects. For two-dimensional fields, activation is shown color-coded, where blue colors denote subtreshold and yellow colors denote suprathreshold activation. For three-dimensional fields, two-dimensional slices of activation are shown. Neural nodes are denoted by circles that are filled if the node is active and empty if inactive. Excitatory synaptic connections are shown by black lines with arrowheads, inhibitory connections by lines ending in black circles; patterned connections are marked with a star. Steerable neural mappings are denoted by blue diamonds. See text for details.

relation (here, to the left of), and the object which the relation uses as a reference position (the reference object, here blue). The spatial transformation system represents these three elements in dedicated dynamic neural fields, the target field, the relational field, and the reference field, respectively. The target field and reference field are defined over two-dimensional space and receive input from the attention field. Whenever there is a peak in the attention field, one of the fields may be brought into the dynamic regime to form peaks. The two-dimensional relational field represents the relative position of a target object with respect to the reference object. The field is defined such that the reference object would be in the center of the field. The relational field also receives input from the production nodes of all spatial relation concepts (e.g., TO THE LEFT OF, see Figure 1). Coordinate transformations between the absolute spatial positions in the target field and the relative positions in the relational field are based on steerable neural mappings (blue diamonds in Figure 1; Schneegans & Schöner, 2012), which are approximated by convolutions here. The architecture has four such coordinate transforms (blue diamonds, from left to right): the first enables the position of an already existing target object to be transformed into the relational field. The second transforms a peak in the relational field into the reference field. The third is analogous to the second but feeds into the target field. These three coordinate transforms enable the architecture to make inferences on an already established mental model. The third transform also accounts for the creation of new objects in the scene: a peak is induced in the relational field from the spatial template that represents one of the spatial relations. The position in space where the peak forms determines where the new object is going to be placed in space. The fourth transform has a crucial impact on the position where the peak forms in the relational field. It transforms the output of the spatial scene representation field and feeds inhibitorily into the relational field, introducing inhibition in positions that are already occupied by objects in the mental model. Due to this inhibition, peaks induced in the relational field tend to shift further outward, avoiding changes to the already established mental model. This is consistent with the preferred mental models that humans tend to build (Ragni & Knauff, 2013).

Process organization

All of the processes within the architecture evolve autonomously in time based solely on the underlying dynamical system. That is, the architecture does not depend on any al-

gorithms or control inputs from the user in order to successfully build mental models; the user only supplies premises. The architecture organizes the processes based on the principles of behavioral organization (Richter, Sandamirskaya, & Schöner, 2012). Every process is represented by a pair of dynamic neural nodes: *intention nodes* (green circles marked "i" in Figure 1) represent whether a process is currently active and determines its influence on the rest of the architecture; *condition-of-satisfaction (CoS) nodes* (red circles marked "c") represent whether a process has been successfully finished and determine the conditions leading to that finish. The architecture has four processes. The intention node of the *reference process* gives input to the reference field and to all production nodes of color concepts that are tied to the reference role. It thus brings the color—and thereby also the spatial position—of the reference object into the attentional foreground, bringing its spatial position into the reference field. If the reference object is not yet part of the mental model in the scene representation field, it is added to it as well (in the center). The CoS of the reference process is a peak in the reference field. The intention node of the *target process* is analogous to that of the reference process. The CoS of the target process is a peak in the target CoS field, which checks whether the target object is represented as part of the mental model. The *relational process* gives input to all production nodes of spatial relation concepts and to the relational field. This induces a peak in the relational field, establishing the position at which the new target object is placed. Lastly, the *reset process* only has an intention node, which inhibits large parts of the architecture in order to remove any self-sustained peaks or turn off self-sustained nodes. This is required before a new premise is supplied to the architecture to prevent activation from previous premises from interfering. Two *precondition nodes* (black circles in Figure 1) ensure that processes are organized in a sequence. They inhibit the intention nodes of the target process and the reset process, respectively. They are in turn inhibited by the CoS nodes of the reference process as well as the target and relational process, respectively. This structure leads to the sequence: reference process, target process, and reset process. The relational process can be active independently of the reference and target process. After the reset process, the architecture is again in a state where a new premise can be supplied.

Results

This section demonstrates how the activation in the architecture evolves as it incrementally builds a mental model of the following four premises: 1. There is a cyan object above a green object. 2. There is a red object to the left of the green object. 3. There is a blue object to the right of the red object. 4. There is an orange object to the left of the blue object. This example shows that the architecture is able to interpret multiple colors and spatial relations, that it can use different target and reference objects across premises, and that it can deal with both determinate and indeterminate cases.

Figure 2 shows activation snapshots of relevant parts of the architecture at six moments in time ($t_1, \ldots, t_6$) during the task. Time points t_1, t_2, and t_6 show the result after supplying and building the mental model according to the first, second, and fourth premise, respectively. Time points $t_3, \ldots, t_5$ show the detailed processes within the architecture that extend the mental model for the third premise.

Determinate cases

At the beginning of the task, the user supplies the first *determinate* premise by activating the three memory nodes for "target: cyan", "spatial relation: above", and "reference: green" (leftmost column, topmost two rows in Figure 2). At t_1, the reference process has already brought the reference object into the center of the reference field (fourth row) and the scene representation field (last row) and has turned off. The relational process is still active but has already induced a peak in the relational field (fifth row) from the template of the spatial relation ABOVE, yielding the new position of the cyan target object. From there, it is represented in the target field (sixth row), the attention field (seventh row), bound there with the CYAN color from the color attention field (third row), and memorized in the scene representation field. After t_1, the reset process is activated, removing all sustained peaks from the architecture except for the ones in the scene representation field.

At time t_2, the mental model has been extended to represent the second premise "the red object to the left of the green object" (second column, last row). Since the second premise is also a determinate case, the processes are analogous to those of the first premise.

Indeterminate cases

The processes regarding the third premise, "blue to the right of red", are shown in more detail at three time points $t_3, \ldots, t_5$. The premise is an *indeterminate* problem because the green object already occupies the spatial position directly to the right of the red object. Data by Ragni and Knauff (2013) suggests that most subjects would place the blue object in the first free position to the right of the red object—to the right of the green object. The model captures this behavior. Shortly before t_3 (third column), the premise is supplied by the user, who activates the memory nodes for "reference: red", "target: blue" (first row), and "spatial relation: to the right of" (second row). At t_3, the reference process brings the color red into the attentional foreground in the color attention field (third row). This brings the red object in the mental model (last row) into the attention field (second to last row) and establishes it as the reference object in the reference field (fourth row). At t_4 (fourth column), the relational process has already activated the production node of the spatial relation TO THE RIGHT OF, which projects into the relational field (fifth row). Crucially, this field also receives inhibitory input from the spatial scene representation field, reflecting the spatial positions of all objects that are already part of the mental model (round blue shapes in the plot). This inhibition leads to the position of

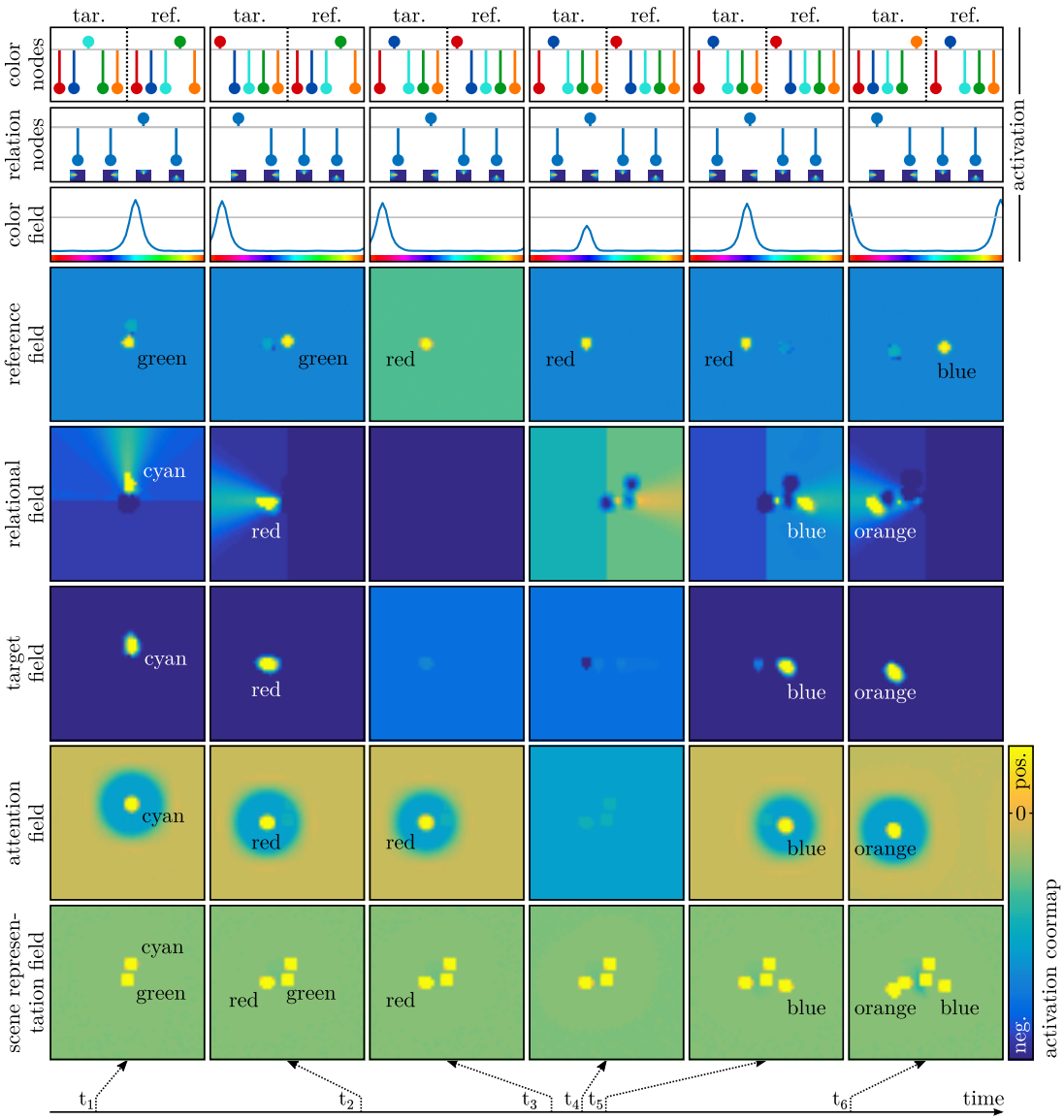

Figure 2: Activation of relevant parts of the architecture as it builds a mental model of four premises about five objects (see text). Each column shows the activation of the architecture at a point in the continuous time shown on the bottom. Activation of memory nodes (first two rows) and one-dimensional fields are plotted on the vertical axis; the threshold of zero is denoted by gray lines. The colors of the circles and bars in the first row correspond to the color concepts the nodes represent. The concepts of spatial relations (second row; LEFT, RIGHT, ABOVE, BELOW) are denoted by icons underneath the bars. All other activation plots are color-coded (color-map, bottom right). The three-dimensional attention field and scene representation field are shown projected onto 2D space by a maximum operation. Peaks in fields are labeled with the colors of the objects they represent.

the target object to be established in the next free position to the right of the red reference object. At t_5 (fifth column), the relational field (fifth row) has formed a peak to the right of the local inhibition, determining the position of the new blue object. The target process brings the peak into the target field (sixth row) and from there into the attention field (second to last row). In this field, the input from the target field overlaps with input from the color attention field (third row), binding the spatial position of the new target object to the color BLUE. From there, the peak comes up in the scene representation field (last row) as well as in the spatial scene representation field.

The next time point, t_6, shows the fourth premise, "the orange object to the left of the blue object"—also an indeterminate problem. This time, two objects (green and red) occupy the positions to the left of the blue object. The relational field (fifth row) shows a peak as well as the inhibitory influence of the two objects that brought up the peak in the first free position to the left of the reference object. The position of the peak is transformed into the target field (sixth row), bound

with the color ORANGE in the attention field (second to last row), and integrated into the mental model in the scene representation field (last row).

Inferences on mental models

Once a mental model has been constructed, the architecture supports inferences. For instance, it is able to extract the spatial relation between objects, even if their relation has not been specified by the premises before. This is enabled by the transformation (leftmost diamond) that induces a peak in the relational field from a given target and reference position. The peak activates the most fitting spatial relation based on its template. Moreover, given a phrase that contains only one object (reference or target) and a spatial relation, the architecture can infer the second object.

Discussion

We have introduced an architecture that captures the neural dynamic processes of building mental models. With only minimal changes, the architecture is based on previous work in dynamic field theory (DFT) that accounts for the grounding of language in actual perception (Richter et al., 2014; Richter et al., 2017). New over the previous architecture is how object representations are instantiated without perceptual input. In particular, we show how the architecture deals with the problem of placing objects in space. The processes of the architecture on indeterminate problems matches data from human subjects, who tend to build a preferred mental model that minimizes change (Ragni & Knauff, 2013). This behavior emerges from the neural dynamics, specifically from the interplay between the already established mental model and the way positions for new objects are established. We furthermore show that the architecture iteratively builds a mental model as multiple premises are supplied in a sequence.

Our approach shares concepts with the information processing account of Ragni and Knauff (2013), in which symbolic elements are placed onto a two-dimensional grid-canvas. Our architecture reframes these operations as neural processes. This includes the emergence of discrete operational stages of processing from an underlying time-continuous neural dynamics (Richter et al., 2012).

Overall, we show that mental models can be captured by the same neural mechanisms that also support the perceptual grounding of language. Experimental signatures of the proposed mechanisms may be sought by asking participants to graphically represent their mental map. Future versions of the architecture may incorporate additional constraints that shed light on how strategies other than the preferred mental model may come about. Extensions of the architecture may be able to account for how abstract relations are represented (Knauff, 2013) by mapping them onto the spatial canvas described here. Establishing mappings between different types of relations may ultimately lead to a neurally plausible architecture of general relational reasoning.

References

Amari, S.-i. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, *27*(2), 77–87.

Barsalou, L. W. (2009). Simulation, situated conceptualization, and prediction. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*(1521), 1281–9.

Gallese, V. & Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology*, *22*(3-4), 455–479.

Johnson-Laird, P. N. (2010). Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, *107*(43), 18243–18250.

Knauff, M. (2013). *Space to Reason*. Cambridge, MA: MIT Press.

Kosslyn, S. M., Ganis, G., & Thompson, W. L. (2001). Neural foundations of imagery. *Nature Reviews Neuroscience*, *2*(9), 635–642.

Logan, G. D. & Sadler, D. D. (1996). A computational analysis of the apprehension of spatial relations. In P. Bloom, M. Peterson, L. Nadel, & M. Garrett (Eds.), *Language and Space* (Chap. 13, pp. 493–529). Cambridge, MA, USA: MIT Press.

Mani, K. & Johnson-Laird, P. N. (1982). The mental representation of spatial descriptions. *Memory & Cognition*, *10*(2), 181–187.

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, *6*(July), 576–582.

Ragni, M. & Knauff, M. (2013). A theory and a computational model of spatial reasoning with preferred mental models. *Psychological Review*, *120*(3), 561–588.

Richter, M., Lins, J., Schneegans, S., Sandamirskaya, Y., & Schöner, G. (2014). Autonomous neural dynamics to test hypotheses in a model of spatial language. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *CogSci 2014* (pp. 2847–2852). Austin, TX: Cognitive Science Society.

Richter, M., Lins, J., & Schöner, G. (2017). A neural dynamic model generates descriptions of object-oriented actions. *Topics in Cognitive Science*, *9*(1), 35–47.

Richter, M., Sandamirskaya, Y., & Schöner, G. (2012). A robotic architecture for action selection and behavioral organization inspired by human cognition. In *IROS 2012* (pp. 2457–2464). New York, NY: IEEE.

Schneegans, S. & Schöner, G. (2012). A neural mechanism for coordinate transformation predicts pre-saccadic remapping. *Biological Cybernetics*, *106*(2), 89–109.

Schöner, G., Spencer, J. P., & the DFT Research Group. (2015). *Dynamic Thinking: A Primer on Dynamic Field Theory*. New York: Oxford University Press.

Zwaan, R. A. (2014). Embodiment and language comprehension: Reframing the discussion. *Trends in Cognitive Sciences*, *18*(5), 229–234.