

To appear in: Houdé, O. & Borst, G. (Eds.), *The Cambridge Handbook of Cognitive Development*, Volume 3: Education and school-learning domains. Cambridge, UK: Cambridge University Press.

Computational methods in education: Neurocomputational models of cognition versus technology as a tool for supporting learning and teaching

Michael S. C. Thomas, Centre for Educational Neuroscience, Birkbeck, University of London, UK (m.thomas@bbk.ac.uk)

Kaska Porayska-Pomsta, Centre for Educational Neuroscience, UCL Institute of Education, UK (k.porayska-pomsta@ucl.ac.uk)

In this chapter, we consider computational approaches to understanding learning and teaching. We consider the utility of computational methods in two senses, which we address in separate sections. In the first part, we consider the use of computers to build *models of cognition*, focusing on the one hand on how they allow us to understand the developmental origins of behaviour and the role of experience in shaping behaviour, and on the other hand how a particular type of model – artificial neural networks – can uncover the way in which the constraints of brain function likely shape the properties of our cognitive systems. In the second half of the chapter, we consider the use of computers as *tools to aid teaching*, in particular in the use of artificial intelligence in education.

These two approaches naturally cross-fertilise. The origin of computational devices in the early 20th century lay in an endeavour to build machines that thought as humans did; in order to have a good computer tool to help teachers, the design of the tool needs to be informed by how children learn. One of the goals of the latter part of the chapter is to provide a basis for informed discussion of whether and how the developmental cognitive neuroscience and artificial intelligence approaches can guide each other for the benefit of their respective aims, and whether artificial intelligence may be able to act as a bridge between developmental cognitive neuroscience research and real-world educational practices.

Part 1. Computers as models of cognition

Humans are biological entities, whatever the sophistication of our cultures and cultural artefacts. When it comes to education, we are, as it were, primates in the classroom. Understanding the operation of the brain – the biological basis of learning – in terms of computation is one perspective of what biological systems do. It is a valuable perspective that helps us understand and unify various properties of the brain – such as the electrical activity of neurons – and how these properties relate to behaviour. However, there may be limitations to the perspective, for example in the properties of biology or the environment

which are de-emphasised or ignored.¹ More widely, the computational perspective fits into the contemporary cultural context of measurement and optimisation familiar in free market societies ('everything has a cost, optimise profits'). Computational modelling is to some extent a method of our time. Those caveats in mind, let us introduce the theoretical framework in which we will consider the use of computational approaches in the first part of this chapter.

The first theoretical framework we will use is *educational neuroscience*. This is an emerging interdisciplinary field that seeks to use new insights into brain mechanisms of learning to inform educational practices (Mareschal, Butterworth, & Tolmie, 2013; Thomas, Mareschal & Dumontheil, 2020). It is not reductionist, in the sense that the field comprises a dialogue between neuroscientists, psychologists, and educators, with an understanding that education is a much broader phenomenon than the changing of brains. Education is intrinsically a cultural, community-based enterprise based on social interaction. Yet to acquire new knowledge and skills, this must be achieved through changing brains. The interaction between neuroscience and education occurs along two main pathways (Thomas, Ansari & Knowland, 2019). The first is an indirect link, where neuroscience findings inform psychological theories, which inform education practices. These might concern specific educational domains, such as literacy or numeracy, or more general aspects of cognition that impact learning, such as executive function skills, emotion, or motivation. The second route is a direct one, where insights from neuroscience help to optimise the brain for learning when the child enters the classroom, such as the impact of diet, sleep, exercise, or stress.

The second framework we will use is *neuroconstructivism* (Mareschal et al., 2007; Westermann et al., 2007). This is a theory of cognitive development that combines a Piagetian constructivist approach – that more complex knowledge and skills are constructed on the basis of simpler knowledge and skills via the child's experience of the world – with a contemporary understanding of functional brain development. The development of functional brain systems is viewed as heavily constrained by multiple interacting factors that are both intrinsic and extrinsic to the developing child. Cognitive development occurs in the context of the constraints operating on the development of the brain that span multiple levels of analysis: from genes and the individual cell to the physical and social environment of the developing child. Neuroconstructivism integrates different views of brain and cognitive development including probabilistic epigenesis (emphasising the interactions between experience and gene expression in shaping development), neural constructivism (focusing on the experience-dependent elaboration of small-scale neural structures), the interactive specialisation view of brain development (stressing the role of interactions between different brain regions in functional brain development), embodiment (highlighting the role of the physical body in cognitive development), Piagetian constructivism (focusing on the child's pro-active acquisition of knowledge), and the role of the social environment for the developing child.

¹ For example, see reservations of Rodney Brooks, a leading artificial intelligence and robotics researcher: <https://www.edge.org/response-detail/25336B>. Brookes argues that planetary orbits around the sun can be described and simulated in computational terms, but no one would argue that planets are computers.

Cognition, computation, education, and the brain

The view of cognition as computation has been a mainstay of cognitive psychology since the 1980s. It leads to research methods that seek to identify mental representations and processes that manipulate those representations. Cognitive psychology has a long relationship with artificial intelligence research, which constructs machines that can operate in intelligent ways. The collaboration of these fields has led to the identification of possible ways that cognition could work, either in humans or in machines. However, there is only one way that cognition *actually* works in humans, and that is constrained by how the brain works. There are things that the brain can do that a conventional (symbolic, rule-based, von Neumann) computer cannot, and vice versa. From a computational perspective, the goal of neuroconstructivism is to identify how the constraints of being implemented in the brain shapes the cognitive processes of the mind. The properties of the brain stem from its biology and its biology is the outcome of a long evolutionary history. This means the way the brain does things may not necessarily be the best, but it will be optimised (by evolutionary selection) given what was available in ancestor species. For example, a biological constraint is that cognition will be performed by neurons. Neural activity produces metabolic waste products which must be cleared away, and changes in neural properties and connectivity require consolidation to be stabilised as robust memories. Together these factors mean that organisms need to sleep – humans are off-line for a third of their lives. On the face of it, this is not an optimal solution for a cognitive system, and it is a limitation that symbolic computers do not suffer from.

What, then, are the *implementation constraints* of performing computation in the brain? The basic unit of computation is the neuron, and knowledge is stored in the strength of the connections between neurons. This means knowledge is built into structure. It means that neural processing systems will be content-specific. The brain, then, is built of a set of **content-specific systems** (be they motor or sensory cortex). It then requires a separate, specialist system whose job it is to **modulate** the activity of the various content-specific systems, to make sure that the appropriate parts are activated and inhibited based on the current context and goals (the role of the pre-frontal cortex). The content-specific systems must be linked by **translators**, and their content integrated by **hubs** (such as the hippocampus for episodic memory, or anterior temporal lobe for semantics). The bread and butter of the brain are its sensory and motor systems. These content systems are **hierarchical**, a sequence of layers each picking up increasingly higher order invariances (conceptual structure) in the information to which they are exposed, from immediate low-level motor actions to long-term high-level plans in the motor system, and from low-level perceptual features to high-level objects in sensory systems. Activity travels simultaneously up and down these hierarchical systems so that expectations (e.g., of the object you will see) can influence low-level processing. The brain exists to serve the body, and there are brain structures dedicated to the evolutionary goals of the organism (eating, sleeping, detecting threats, bonding, mating, fighting). The **emotion** (limbic) system interacts with the modulatory system to influence its goals; it influences regional properties of processing in the cortex through altering **neurotransmitter levels** (e.g., to alter arousal); and it **conditions the body** to be in the appropriate state for the current situation (e.g., fight or flight responses).

With respect to education, there is a **many-to-one** mapping between content-specific systems of the brain and concepts utilised in psychology. So for example, 'addition' in mathematics class involves multiple representations of knowledge in different brain systems (visual symbols, representations of quantity), motor sequences (of pencil movements), and strategies (retrieval, execution of procedures), in a complex sequence of activity over time, and sometimes involving iterations of physical interaction with the environment (move head and eyes to look at problem, write with pencil, look back to problem). "Learning" as an educational concept involves the **on-going interaction between perhaps eight different neural systems**, including **reward-based** processing systems and a system involved in the **automatization** of movements (Thomas, Ansari & Knowland, 2019). Notably, educational psychology tends to focus on the acquisition of abstract knowledge underpinned by cortical mechanism (e.g., to learn multiplication, the system must link language-encoded times tables to procedural knowledge for linking number symbols, and to the semantic underpinning of quantity). However, from the biological perspective of the social primate sitting in the classroom, this function is probably only the brain's fourth most important priority. Before it come, respectively, movement, emotion, and social relations (e.g., leaning on the desk and fiddling with your pencil, feeling anxious about maths, wanting to whisper a question to your friend to find out why Sienna doesn't like you anymore). Learning is optimised when the first three are aligned in the service of the fourth (e.g., motor activities are relevant to the topic, there is excitement and curiosity for learning, and learning is supported by the peer group).

There are multiple methodological approaches to investigate this complex interactive system within developmental cognitive neuroscience. Computational modelling represents a set of formal methods to specify the representations and processes involved in various components of the cognitive system. Computational methods are widely used in other scientific fields to simulate the behaviour of complex systems, such as in meteorology or astrophysics. Formal models have certain virtues. For example, they enforce precision on sometimes vague implicit or verbal accounts of how systems work; if a theory is implemented as a working system, it can test the viability of the theory to produce the observed behaviour it claims to explain; and models able to unify diverse phenomena provide parsimony. Computational models are particularly important in studying systems with multiple interacting components, where the behaviour of the whole system emerges through complex interactions. Once a model is constructed, it can be applied to new situations, and generate novel testable predictions, for example, when its parameters are set to atypical values (e.g., as we'll see later, to capture disorders such as dyslexia or Attention Deficit Hyperactivity Disorder for models of reading and decision-making, respectively).

The main disadvantage of models is that, by definition, they require simplification. As Box and Draper (1986) say, 'all models are wrong, some are useful'. This both poses the challenge of ensuring only irrelevant details are simplified away in building a model, and also finding a balance between building a model complex that is enough to capture the target phenomenon but not too complex so that the model's own functioning cannot be understood (Lewandowsky, 1993; McCloskey, 1991). The ultimate goal of modelling, after all, is to progress theoretical understanding.

Computational models of cognition have used different types of computational formalism. Some of them rely on explicit rules for encoding knowledge (*IF x THEN y*) (e.g., Ritter, Tehranchi & Oury, 2018). Some employ formalisms from probability theory, where cognition is viewed as updating a probabilistic understanding of the state of the world in the light of new data (the Bayesian approach; Gopnik & Bonawitz, 2015). Models that focus on the computations that can be performed by neural systems can differ depending on whether they focus on the temporal dynamics of the system (dynamical systems theory; Spencer, Perone & Buss, 2011); or the information encoded in representations (connectionism or artificial neural network models; Thomas & McClelland, 2008; Spencer, McClelland & Thomas, 2009). In the following section, we focus on artificial neural network models, applied to education-relevant cognitive abilities. We do so, because these machine-learning systems have the attractive property of learning their knowledge representations by exposure to a structured learning environment; they are therefore ideally suited to studying learning, development, and mechanisms of change (Mareschal & Thomas, 2007).

Artificial neural network models of education-relevant cognitive abilities

Artificial neural network models (henceforth ANNs) have been applied to modelling a range of phenomena in cognitive development, from sensori-motor processing in infancy (e.g., object recognition), routine motor sequences, categorisation, aspects of language such as vocabulary, morphology, and syntax, and reasoning on Piagetian problems (Botvinick & Plaut, 2004; Elman et al., 1996; Mareschal & Thomas, 2007; Shultz, 2003). Models target development in particular cognitive domains and for restricted behavioural phenomena (e.g., the ability to sort rods of different lengths into serial order; Mareschal & Shultz, 1999). The approach is therefore an analytical one, pulling cognition apart into component parts, and is as a consequence reliant on theories of developmental cognitive neuroscience to identify the relevant components.

The parts of an ANN are as follows. The basic elements are *simple processing units* with activity levels, analogous to neurons and electrical neural firing rates. A unit's activity level alters the activity levels of other units to which it is connected, based on the strength of the *connections* between them. The connections are analogous to axons, synapses, neurotransmitters and dendrites. Units have an *activation function* that determines how much they will alter their activation level depending on the level of stimulation (excitation, inhibition) they are receiving from other units. Units are typically organised into *layers*. A layer represents information through a pattern of activations across its units. In *neural* models, models are tested against their ability to simulate neural activity. In *cognitive* models, representations correspond to concepts and models are tested against their ability to simulate behaviour.

Layers are usually defined as inputs, outputs, and intermediate layers that facilitate the mapping between inputs and outputs. The layers and pathways in a model are referred to as the *architecture* (e.g., the architecture of a model of the reading system is shown in Box 1). An untrained ANN has small random connection weights. The ANN is exposed to a *structured learning environment* (or training set), which specifies the sets of input-output mappings it must learn. Input-output pairs are presented to the network, and a *learning*

algorithm is used to adjust the connection strengths so that the network gradually learns all the input-output pairs through multiple exposures to the mappings. There are a variety of learning algorithms, which generally serve to alter the network connections to optimise some function, be it the accuracy of input-outputs, the conciseness of a set of representations, or how accurately the network can predict the reward gained by a particular action. In *agent-based* models, the system is an agent whose actions alter the subsequent experience of the environment (for more detail, see Elman et al., 1996; McLeod, Plunkett & Rolls, 1998; Thomas & McClelland, 2008). These components are summarised in Table 1, left-hand column.

Table 1. Components of two different types of model, those used to simulate cognitive mechanisms, and those used as theoretically informed artificial intelligence tools to support learning and teaching

<i>Cognitive model components</i>	<i>Artificial Intelligence in Education model components</i>
Stipulation of Theoretical domain of relevance – what is to be modelled (e.g., reading development) and what is to be simplified (e.g., vision and audition)	Domain model responsible for representing the knowledge and related operations that are the object of learning (e.g., maths)
Architecture of model specifying inputs, outputs, pathways, internal layers, parameter settings (e.g., pathways linking orthographic, phonological and semantic representations)	Model of the learner which represents what the learner knows at any given point as well as their emotions and motivational states
Learning algorithm specifying how structure and parameters of the model will change based on training experiences or development	Model of pedagogy taking into account the domain to be mastered and the pedagogical strategies and tactics that are appropriate in that domain
Representational format for inputs and outputs (e.g., code for speech sounds, code for written letters)	Communication model offering strategies for how to realise any given pedagogical strategy
Specification of Structured learning environment – frequency and nature of experiences (e.g., associations between written and spoken forms of words); in agent-based modelling, the agent’s actions determine the next input from its environment, which may also contain other agents	

ANNs have been applied to a number of cognitive models relevant to education. Perhaps the most attention has been paid to capturing the development of **reading** (e.g., Seidenberg & McClelland, 1989; Harm & Seidenberg, 2004; Plaut et al., 1996; see Box 1 for an example model). These models focus on content-specific pathways which learn to translate between

structured representations of a word's written form (orthography), its spoken form (phonology), and its meaning (semantics). The model is exposed to a learning environment in which it is presented with instances of associations between the written and spoken form of words, encountering them with a frequency based on the occurrence of these words in naturalistic corpora. The accuracy of the model in reading depends on how often it encounters words, but also on the complexity of the relationship between written and spoken forms, which may be fairly transparent (e.g., Italian) or complex (e.g., English). Box 1 provides an example of some of the implementation details of a specific model, and how models have been extended and tested by brain imaging data.

Representations of **meaning** in cognitive models are usually depicted in terms of sets of semantic features that define a concept. More recent models have begun to capture semantic representations in terms of a hub, where information from diverse modalities of a concept can be unified (e.g., sound, touch, visual features, smell, movement, verbal descriptions) (e.g., Chen, Lambon Ralph & Rogers, 2017). Meaning has been represented as sequences of associated concepts over time structured into events or episodes (Hoffman, McClelland & Lambon Ralph, 2018; Elman & McRae, 2017). And access to semantics has been proposed to require external modulatory control processes activating and inhibiting content representations (Hoffman, McClelland & Lambon Ralph, 2018).

Such models acquire generalised representations of meaning, gradually extracting patterns over multiple exposures to individual instances of, say, *dogs* or *cars*. However, the brain also has a structure, called the hippocampus, for snapshot learning of individual **episodic memories**, e.g., where and when you saw a specific dog. Knowledge of individual episodes must somehow be transferred to the cortical representations of general **semantic knowledge**. This process has been studied in models of complementary learning systems, where the hippocampus supports **consolidation** of knowledge in the cortex, partly by replaying memories during sleep (McClelland, McNaughton & O'Reilly, 1995; O'Reilly et al., 2014).

Models of **numerical cognition** have focused on capturing basic tasks such as number comparison and simple addition, since developmentally, more complex tasks such as multidigit arithmetic and symbolic mathematical reasoning build on these simpler tasks (Zorzi, Stoianov & Umiltà, 2005). In these types of models, the acquisition of number concepts involves the mapping between an analogue code of quantity, representations of number symbols (e.g., Arabic numerals) and verbal numerical expressions (e.g., Dehaene & Cohen, 1995; Campbell, 1994). Such models have attempted to account for phenomena such as the distance effect (that is it is easier to select the larger of two numbers when they are far apart than when they are close) and the size effect (that for a given distance, it is easier to compare small numbers than large numbers) (see Dehaene, 2003, for review). Models of simple arithmetic aim to address the fact that competent adults can use a combination of fact retrieval from memory and procedures for transforming the problem if memory search fails, and therefore must combine multiple pathways (Zorzi, Stoianov & Umiltà, 2005).

Models have considered **executive functions**, for example in cognitive control (Botvinick et al., 2001), in short-term memory (Haarmann & Usher, 2001), and even switching between

the bilingual's two languages (Filippi, Karaminis & Thomas, 2014). These models include modulatory mechanisms that influence or retain activation states in the content-specific systems to which they are connected. However, control of behaviour is also sometimes construed within a reinforcement learning framework, where decisions about behavioural choices depend on a history of the rewards received for different actions. Models of **reward-based decision making** have been influenced by a growing understanding of the role of the dopamine neurotransmitter system in the striatum, where neural activity has been found to follow the accuracy of the individual's predictions of the rewards they will receive for their actions (Zeigler et al., 2016). These models have been extended to consider the possible origins of impulsivity in Attention Deficit Hyperactivity Disorder, construed in terms of changes to the model's initial computational parameter settings, for example in the weight given to small short-term versus larger longer-term rewards (Zeigler et al., 2016, for a review).

Findings from ANN models of the acquisition of education-relevant abilities point to the importance of the quality of the representations for driving the learning of more complex abilities (e.g., phonology for reading, an analogue code of quantity for numeracy); the importance of sufficient capacity and plasticity in processing systems to acquire target skills; the importance of context-appropriate control of the activation states in content systems, and the importance of representative exposure to the problem domain.

Finally, low-level sensory systems do not tend to be the focus of education-relevant computational models. Nevertheless, they can sometime be relevant because education is seeking to shape brain systems that have evolved for other purposes (so-called **neuronal recycling**; Dehaene, 2005). The computational constraints of these brain systems may influence behavioural patterns as new culturally determined skills are acquired. For example, in some scripts, written letters can be mirror reversals or rotations of each other (e.g., b, d, p, q in English). The visual system develops to recognise objects irrespective of their orientation, a constraint that must be overridden to separate these orientation-specific letters. The result is initial characteristic errors of confusion of these letters (and numbers such as 2 and 5) (see, e.g., Blackburne et al., 2014). Recently, ANN models have been successfully applied to capturing the development of visual object recognition. Advances in deep neural networks have enabled computer scientists to produce much more powerful systems for recognising complex objects within visual scenes. These models have multiple layers, with each higher layer extracting more complex features from the visual input. Deep networks are very powerful learning systems but very specific to the content on which they are trained. Two points are notable. First, the types of representations developed in the sequence of layers in the artificial neural networks appear to capture the types of representations found in the hierarchy of processing areas in the ventral (object recognition) stream of the human cortex, validating deep learning as a useful perspective on brain function (e.g., Rajalingham et al., 2018). Second, the similarity of the ANN's representations to the brain's representations depends on how many layers the model has. Beyond a certain number of layers, the ANN's performance begins to *exceed* human accuracy on image classification, and the layers' representations *cease to be humanlike* (Storrs et al., 2017).

Box 1. Example of an ANN model of reading development

A great deal of research has focused on developmental models of reading. Initial models addressed how ANNs could learn the mapping between orthography and phonology by repeated exposure to a word's written and spoken forms, how such models could accommodate both regularity in these mappings (*mint*, *hint*, *tint*) but also exceptions (*pint*), and how they could extract the general function linking spoken and written forms to enable them to read aloud nonwords (e.g., *gint*) (e.g., Seidenberg & McClelland, 1989). Subsequently, models extended to consider the possibility that a written word's meaning could be retrieved either by a direct mapping from orthography to semantics, or by generating its spoken form and using this to access semantics. Similarly, the spoken form could be retrieved either directly from the written form, or the written form could be used to retrieve the meaning which could then be used to retrieve the spoken form (Harm & Seidenberg, 2004; Plaut et al., 1996). In this way, the reading system therefore has multiple pathways, and there may indeed be a division of labour between them. For example, it might be more efficient for a system to learn regular mappings via the direct orthography to phonology route, and the exceptions (like *pint*) via the semantic route.

Figure 1 shows the architecture of one implementation of the multiple pathway architecture (grey elements depict unimplemented input and output systems) (Harm & Seidenberg, 2004). Notably, many of the connections between layers of units are bi-directional, so that activation can flow around the network. In the model, semantics was represented over 1989 units, orthography over 225 units, and phonology over 200 units. Intermediate layers helped learn the mappings between these codes. The size of the intermediate layers was determined merely by what worked, or as the authors put it in one case, 'the number 500 was chosen from pilot studies; it is a number large enough to perform the mapping without being too computationally burdensome' (p.677). The model was trained on 6,103 monosyllabic English words, consisting of all monosyllabic words and their most common inflections. The pathways were trained separately, using an algorithm based on backpropagation through time. This algorithm adjusts connection weights to reduce output errors and also accommodates cycling activation. The model was trained for around 700,000 word presentations, first learning an oral language system, then linking written forms to it (see Harm & Seidenberg, 2004, for full details). One notable finding of this model was that, given its multiple pathways, the system initially learned to retrieve meanings from written forms through accessing phonology, because this mapping is largely regular, and the pathway from phonology to semantics is already established. But gradually, the system learned the more complex direct mapping from orthography to semantics, which delivers faster reading.

The dynamics predicted by the computational model were subsequently testable by advances in brain imaging. Dynamical causal modelling techniques applied to functional magnetic resonance imaging (fMRI) data were able to reveal which brain regions dynamically drive which other brain regions during reading (Richardson et al., 2011). Figure 2 shows the strength of the dynamic modulation of the activity between connected regions during a reading task, respectively for low-level vision, visual word processing, phonology and semantics. It reveals that orthography and phonology interact with each other during reading, and both drive semantics; but notably, early visual areas directly drive activation in

both orthographic and phonological areas, with phonology also showing some indication of top down modulation of low-level visual areas. These data confirm the multiple routes, hierarchical nature, and interactivity of the reading system.

Subsequent computational models of the reading system have sought to include further constraints from neuroanatomy (e.g., a dorsal route linking auditory perception directly to motor output, for repeating words without retrieving their meaning; Ueno et al., 2011); have considered how alterations in the computational properties of the system, either in the number of units in the mapping pathways or the quality of the phonological representations, could produce developmental trajectories resembling dyslexia (e.g., Harm & Seidenberg, 1999); how certain kinds of behavioural interventions may ameliorate the developmental deficits (Harm, McCandliss & Seidenberg, 2013; Thomas et al., 2019); and how variation in network parameters may produce individual differences in development and potentially provide a mechanistic link to both genetic levels and environmental variables such as socio-economic status (SES; Thomas, Forrester & Ronald, 2013, 2016). In as much as cognition is viewed as computation, genetic effects must unpack as modulation of neurocomputational parameters; and correlates of SES include variations in the growth of brain structures and variation in the level of cognitive stimulation.

Figure 1. Architecture of the Harm and Seidenberg (2004) model of reading, showing the specified representations (semantics, phonology, orthography), pathways, and directions of activation flow between them. Greyed elements (motor, auditory, visual systems, context) were unimplemented, but assumed, components.

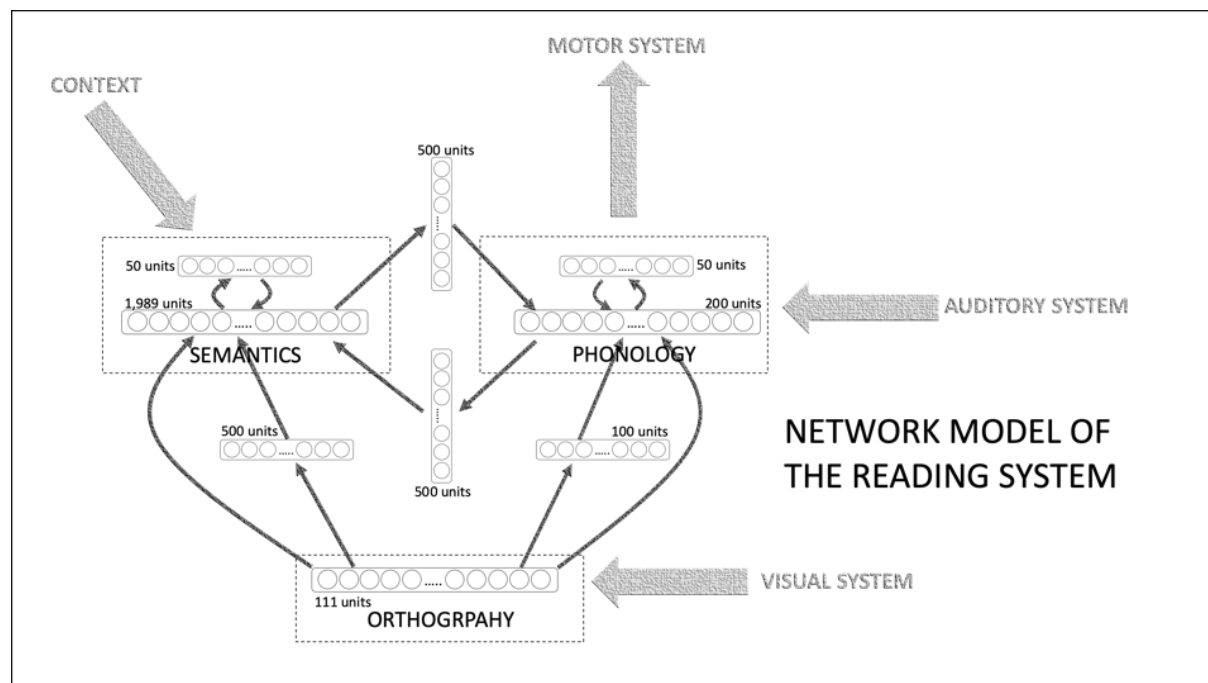
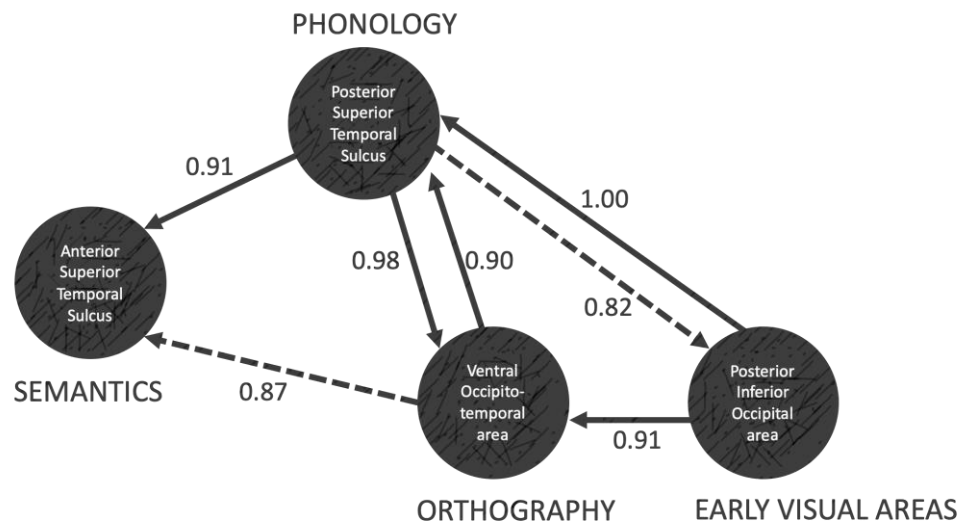


Figure 2. Summary of dynamic causal modelling of functional magnetic resonance imaging data, showing which brain regions involved in reading appeared to causally drive other regions (Richardson et al., 2011). Values show probabilities for modulatory connections

during the reading task. Connections above threshold are indicated by solid black arrows. Strong trends are indicated by black dashed arrows.



Part 2. Technology as a tool for supporting learning and teaching

The study of human learning using computational modelling dates back at least to the 1970s and the formal advent of what is now known as Artificial Intelligence in Education – a research field that lies at the intersection of the broader studies in Artificial Intelligence and the Learning Sciences (e.g., Woolf, 2007). Given that computational modelling of learners in context represents a defining characteristic of AIED technologies (with Intelligent Tutoring Systems providing one example of such technologies), there is a natural overlap between the preceding computational modelling approaches used in educational neuroscience (henceforth EdN) research, and those used in AIED systems.

However, there are also some fundamental differences between the primary motivation and goals of AIED and EdN. These differences relate specifically to the emphasis that each field puts on the importance of neuro-cognitive vs behavioural fidelity of its models, as well as their respective reliance on access to and their immediate application in pedagogical practices at the front-line. In particular, while the primary goal of EdN is to gain fundamental understanding of neural processes related to learning in order to inform a general theory of how the brain works, AIED is concerned with creating environments which form an explicit part of educational interventions from their inception to their delivery in real-world educational contexts. In other words, in AIED systems, neurocognitive fidelity of the models is a highly desirable but not a *necessary* condition for their successful implementation in educational practices.

In the second part of this chapter, we briefly introduce the AIED perspective as an important area in which computational modelling of learning and teaching behaviours forms a central part. Our goal is to provide a basis for informed discussion of whether and how the AIED and EdN can guide each other for the benefit of their respective aims, and of the extent to which AIED may be able to act as a bridge between EdN research and real-world educational practices.

Artificial Intelligence in Education (AIED)

AIED is a subfield of Artificial Intelligence (AI) and the Learning Sciences (LS), which seeks both to understand the behavioural correlates of learning and teaching processes, and to computationally model individual learners as they engage in learning of a particular subject domain in real-time. While utilising AI's techniques and feeding into the Learning Sciences theories and practices, such modelling is essential to enabling educational software environments to provide adaptive, in-the-moment learning and teaching support, e.g. pedagogical feedback to learners, or advice to teachers on how to support individual learners in context. AIED research investigates: (i) how meaningful interactions between teachers and learners develop; (ii) what factors in the physical learning environment contribute to successful learning (be-it software environment, or a combination of software and a broader context in which learning takes place); and (iii) what kind of pedagogical feedback may be more or less conducive to learning by particular types of learners within a specific learning domain and circumstances (e.g., Porayska-Pomsta & Bernardini, 2013; Woolf, 2007).

There exists a whole plethora of different forms of AIEd technologies, from **Intelligent Tutoring Systems** that focus on supporting mastery learning in one-on-one learning contexts (e.g., Cognitive Tutors – Corbett et al., 1997) to **collaborative learning environments** that support learning interactions amongst groups of learners (e.g., Cukurova et al., 2018). Regardless of their specific application, the key common characteristic to all AIEd environments is that their functionality is underpinned by mutually informing set of modelling components, including: (i) a *domain model* which is responsible for representing the knowledge and related operations that are the object of learning (e.g., maths); (ii) a *model of the learner* which represents what the learner knows at any given point as well as their emotions and motivational states; (iii) a *model of pedagogy* which takes into account both the domain to be mastered and the pedagogical strategies and tactics that are appropriate in that domain; and (iv) a *communication model*, which offers strategies for how to realise any given pedagogical strategy. The exact way in which these different components will be implemented and utilised in any given system will depend on the context in which they are to be deployed, the specific intervention goals, hardware employed (which may determine what user behaviours can be detected and modelled feasibly in real-time), and data available. However, regardless of their exact implementations, a learner model is generally considered the essential component of any environment that aims to adapt its pedagogy and interaction to individual learners. These components are included in Table 1 (right column), and contrasted with those employed in cognitive models of learning and development.

Unlike the computational modelling employed in EdN which tends to rely on neurocomputational approaches, AIEd environments are *a priori* agnostic with respect to the type of AI that underpins their models. As such, AIEd technologies employ a diverse range of AI techniques from the so-called **good old-fashioned rule-based AI** (GOF AI) to machine learning (ML). GOF AI requires explicit representation of knowledge, which reflects an ontological conceptualisation of the world and actions that are possible therein, along with some well-defined measures of success in terms of concrete goals and goal satisfaction constraints. For example, in the context of maths tutoring, the ontological representations will relate to the specific sub-domains of maths, say – misconceptions in column subtraction, and rules that define the possible operations on the given subdomain. The goal satisfaction in this case may be in terms of student’s correct or incorrect answers. The rules are typically elicited through questioning of human experts in a given domain, by observing their expertise in real contexts or by hand annotating data (video recordings, interaction logs, etc.) of humans engaging in specific tasks of interest. By contrast, **machine learning** (ML), such as implemented by neural networks, learns solutions from first principles by applying statistical classification methods to large data sets, as discussed in the first part of this chapter. Both of these broad approaches have their strengths and weaknesses in terms of the extent to which they lend themselves as a basis for enhancing our theoretical understanding of learning and teaching processes, or for supporting teaching and learning practices.

Specifically, the key advantage of knowledge-based systems is that they require a detailed understanding of the domain, in order for knowledge ontologies to be constructed, thus also potentially leading to a greater understanding of the domains represented, and the fact that the resulting ontologies are transparent, inspectable, and often understandable by

humans (Davis, 1993; Russell & Norvig, 1995). For example, in **Cognitive Tutors** (henceforth CTs; Corbett et al., 1997), which were originally created as a testbed for the ACT-R cognitive theory of rational thought and problem solving (Anderson et al., 1990), this transparency is key for delivering fine-grained and tailored moment-by-moment feedback to students, and to performing diagnoses (*sic* learner modelling) of learners' developing knowledge and understanding. Here, learner modelling involves keeping track of (i) the learner's progress through a solution and (ii) the growth of learner's knowledge over time. In CTs, the diagnoses are based on the specification of declarative knowledge, e.g.:

“When both sides of the equation are divided by the same value, the resulting quantities are equal”

and procedural knowledge expressed as production rules that apply to a particular stage in a problem-solving episode, e.g.:

“IF the goal is to solve an equation for variable X and the equation is of the form $aX = b$, THEN divide both sides of the equation by a to isolate X”

(Corbett et al., 1997). The production rules are annotated for correctness and specificity of the solutions that they offer. During problem solving a CT keeps track of all the solution steps committed by the learner and identifies the production rules in its database that correspond to learner's solution steps. The annotations associated with each production rule provide the basis for the assessment of the quality of the learner's steps and their problem-solving strategies, which in turn allows the system to choose appropriate feedback. These decisions can be examined in detail and, if necessary or desired, full traces of the diagnoses performed by the CT can be given back to the teachers or learners as an explanation of the system's assessments and of its choices of pedagogical feedback.

CTs provide but one example of how knowledge can be represented in an AI tutoring system and of how learners' knowledge growth can be modelled and supported using a GOF AI approach. Other successful examples of knowledge and learner modelling include constraint-based models which describe a given subset of a knowledge domain in terms of constraints and constraint satisfaction conditions which can be matched to student actions to guide the system's adaptation of its feedback (Ohlsson & Mitrovic, 2007). Topic networks can be also used to represent specific areas of a subject domain taught, allowing the system and the students the flexibility to choose which topics should be covered next (Beale, 2013). On the other hand, models of learners' emotional and motivational states during learning, often rely on probabilistic approaches (Conati et al. 2018; Porayska-Pomsta & Mellish, 2013; Mavrikis, 2008), with the corresponding Bayesian network representations typically being constructed by hand, based on limited, but fine-grained, observational and interaction data, rather than being machine learned. While offering a relatively high degree of inferential transparency, the disadvantage of knowledge-based systems is that they are cumbersome and time consuming to construct; they are by their very nature limited to small subsets of domains modelled; they may reflect practitioners' theories about their own practices rather than the actual practices; and the data on which they operate may be inaccurate or incomplete, as they rely on directly observable teachers' and learners' behaviours – and these in turn may be difficult to detect and diagnose reliably. Any of these factors

individually or in combination may affect the educational efficacy of the GOF AI based systems.

By contrast, ML carries substantial promise both in terms of reducing the effort required to specify knowledge ontologies and in being able to go beyond the knowledge we have ourselves, and in so doing (the questions of bias and correctness of the base models notwithstanding) – in driving more accurate decision-making than our own capabilities allow for. Here, one of the most exciting aspects of ML is that it can discover new associations in the world and predict future outcomes based on prior data in complex domains which may be otherwise hard for us to grasp and analyse efficiently. Recently, given growing availability and access to voluminous educational data (e.g., from MOOCs and commercial educational apps), these advantages of ML have been seized on in the context of learning analytics and educational data mining research (Baker, 2009; 2010; Macfadyen et al., 2014). As well as being very valuable in shedding light onto various relationships between learner behaviours and learning outcomes, these methods are increasingly used to underpin systems that aid *in situ* pedagogical decisions of teachers, for example through dashboards (Martinez Maldonado et al., 2014), i.e. reporting tools which offer teachers data and metrics related to learners' activities at an individual student or group levels. Given the disadvantage of ML approaches is their lack of inferential transparency and explainability, there is also a growing tendency to combine ML and GOF AI approaches at different stages of system implementation and levels of functionality to compensate for each of those paradigms' limitations (Li et al., 2011). For example, cognitive tutors described previously, utilise ML to learn any new problem-solving strategies employed by students as they interact with the systems, thus increasing their diagnostic flexibility and reducing both the effort and potential inaccuracies that are involved in constructing such systems.

Cognitive Fidelity vs Computational Efficiency

One critical difference between computational models employed in cognitive neuroscience, including EdN, and those used within AI, including in AIEd, is that the key criterion for the success of the latter is not whether they are able to model human brain exactly, but rather whether they can autonomously engage in decision-making, and/or semi-autonomously – in a contingently credible interaction with humans. The goal of such systems is not to replace the human and human decision-making (e.g. as driverless cars might do), but to enhance such decision-making either by offering insights that might be otherwise difficult for the human to gain without the help of technology, or by triggering some desired thinking and behaviour (e.g., Porayska-Pomsta & Rajendran, 2019). Hernandez-Orallo and Vold (2019) refer to the latter function of AI models as *cognitive enhancers*, which they propose can vary in terms of their autonomy and coupling with the human. For example, dashboards that offer learning analytics to teachers may be considered cognitive enhancers that are loosely coupled with humans, since the decisions that are made on the basis of the information given, and indeed whether such information is considered at all, are left entirely to their users. By contrast, Intelligent Tutoring Systems such as Cognitive Tutors, could be considered as relatively tightly coupled enhancers, since their decisions are autonomous and they impact directly on the course of their interactions with the users through accurate learner modelling, while also seeking to optimally compensate or enhance the learners' skills, knowledge, and behaviour.

The emphasis on computational efficiency as opposed cognitive fidelity is a necessary compromise that, arguably, accompanies all AI models, of which AIEd models are a specialised subtype. The tension between cognitive fidelity and computational efficiency was always present in AI developments, leading to two conceptions of AI. The first is a general view of AI where the aim is to replicate humanly thinking and behaviour exactly, and which is presently still considered unattainable. The second is a narrow view where the aim is for an AI agent to act in a sufficiently humanly manner by emulating to some extent rational thinking and behaviour, given a set of known constraints and constraint satisfaction conditions that define the world within which such an agent operates, i.e., in an environment in which thinking/computation can be accomplished (Russell & Norvig, 1995, Davis et al., 1993). The kind of AI systems that are presently developed belong to the narrow AI category, where there is an explicit understanding that the AI models are neither exact replicas of the human brain, nor are they complete. Although from the EdN point of view, this lack of cognitive fidelity or completeness may be considered a limitation, in educational contexts, provided that the models lend themselves to being inspected and modified by the users, these seeming limitations can offer important benefits for learning and teaching. This is because such dependency requires an active effort from the users to engage in completing or correcting those models, which in turn relies on and further develops the users' critical thinking and metacognitive competencies (Bull & Kay, 2016; Conati et al., 2018). The branch of the AIEd research which focuses on this affordance relates to the so-called **Open Learner Models (OLMs)**, with research to date demonstrating how OLMs can be used both as a mirror by the learners to help them improve self-monitoring and self-regulation skills (Azevedo & Aleven, 2013; Long & Aleven, 2013; Porayska-Pomsta & Rajendran, 2019), and as a magnifying glass by educational practitioners who want to gain a better understanding of their learners for the purpose of improving their pedagogy (Martinez Maldonado et al., 2014; Bull & Kay, *ibid*; Porayska-Pomsta, 2016). We consider and exemplify different forms of OLMs further in the following subsection.

Open Learner Models (OLMs)

OLMs are student models, i.e., representations of student cognitive and/or affective states, that allow users to access their content with varying levels of interactivity and control (Bull, 1995; Bull & Kay, 2016). Originally, OLMs have been designed to improve model accuracy by enabling students to adjust the models' diagnoses and predictions, if such were deemed inaccurate by the students. Over time, the use of OLMs revealed substantial educational potential in encouraging learners to self-assess, reflect and ultimately self-regulate, because engagement with such models requires the students to understand and evaluate their own decisions and behaviours. Different types of OLMs include models that are:

- (i) *scrutable*, i.e. users may view the models' current evaluation of relevant student's states and abilities;
- (ii) *cooperative* or *negotiable*, where the user and the system work together to arrive at an assessment of student performance;
- (iii) *editable*: users can change directly the models' assessments and even the underlying knowledge representations at will.

In the cognitive tutors' tradition, Long and Alevan (2016) designed a scrutable OLM to help students to self-assess their knowledge in order to share with the system the responsibility for selecting the next problem to work on. Their system employs similar domain knowledge representation and problem selection mechanism to those employed by the Cognitive Tutors, as described earlier. The system evaluates student's problem-solving steps against a set of example solutions and based on this, using Bayesian Knowledge Tracing, it determines which knowledge components the student needs to learn (Alevan & Koedinger, 2013). The probabilities generated over the knowledge components are visualized for the students in terms of 'skill bars' (Figure 3), which they can compare with their own self-assessments. In this approach, student self-reflection constitutes an explicit learning goal, which has been shown to be key in significantly improving learning outcomes for the students who used this OLM.

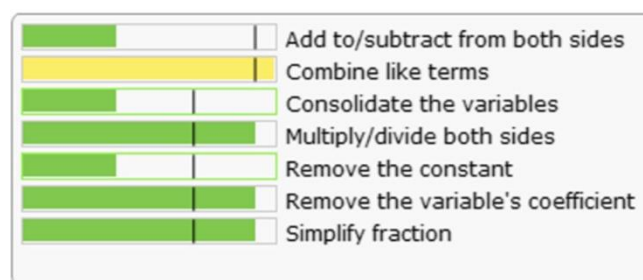


Figure 3. Long and Alevan's (2016) skills meter bars indicating the level of student skill mastery.

An example of a negotiable model is offered by Mabbott and Bull (2006) who created an OLM that allows the learners to 'persuade' the system to change its assessment of their knowledge. To do so, the learners can register their disagreement with the system's assessment and propose a change. At this point, the system will explain why it 'believes' its current assessment to be correct and will provide evidence to support these beliefs, e.g. by showing samples of the learners' previous responses that may indicate a misconception. If the learner still disagrees with the system, they have a chance to 'convince' the system by answering a series of targeted test questions from the system, which keeps a detailed representation of the user on-task interactions and its assessments of the user's understanding given their behavioural patterns and correctness /quality of their solutions.

Basu et al. (2017) designed a fully editable OLM which allows students to construct models of their knowledge by exploring concepts, properties and relations between them in open ended exploratory learning environments. This OLM is underpinned with hierarchical representation of tasks and strategies (implemented as a directed acyclic graph) that may be needed to solve a problem. The system allows for the expression of a particular construct or strategy in multiple variations that relate to each other, which in turn gives the system an ability to assess both desired and suboptimal implementations of a strategy by the learner. Based on this, the system can analyse learners' behaviours by comparing their action sequences and the contexts associated therewith against the strategy variants to offer targeted support when the users seem to flounder. This representation allows for a conceptual support to be given to the user at a fine-grained level of detail, e.g., low-level

objects description in terms of their properties, relations between them and temporal ordering of actions that could be performed on them. In turn, this allows the system to guide the user in editing the model through relatively simple step-by-step interfaces for the different modelling tasks, gradually building users' confidence in their abilities, their buy-in to the system's advice and prompts, ultimately significantly increasing the learning outcomes for the users (Basu et al., 2017).

In the TARDIS system, we implemented a scrutable OLM, in the context of emotional self-regulation in job interview simulations involving AI agents acting as recruiters. TARDIS collects evidence from the simulations, based on low-level signals such as the users gaze patterns, gestures, posture, voice activation, etc., and uses machine learning techniques to predict from this evidence the quality of behaviours known to be important for effective interviews, such as appropriate energy in the voice, fluidity of speech, and gaze management (Porayska-Pomsta et al., 2014). Figure 4 shows how the data are displayed to the user, with the pie charts referring to four qualities of interest such as energy manifested in the users' interactions (which may indicate engagement), fluidity of the interaction (which may be indicative of user confidence), spatial extent which evaluates expansiveness of gestures (these may need to be controlled during a job interview) and overall activation (i.e., users initiations of interaction and responses to agents' initiations).

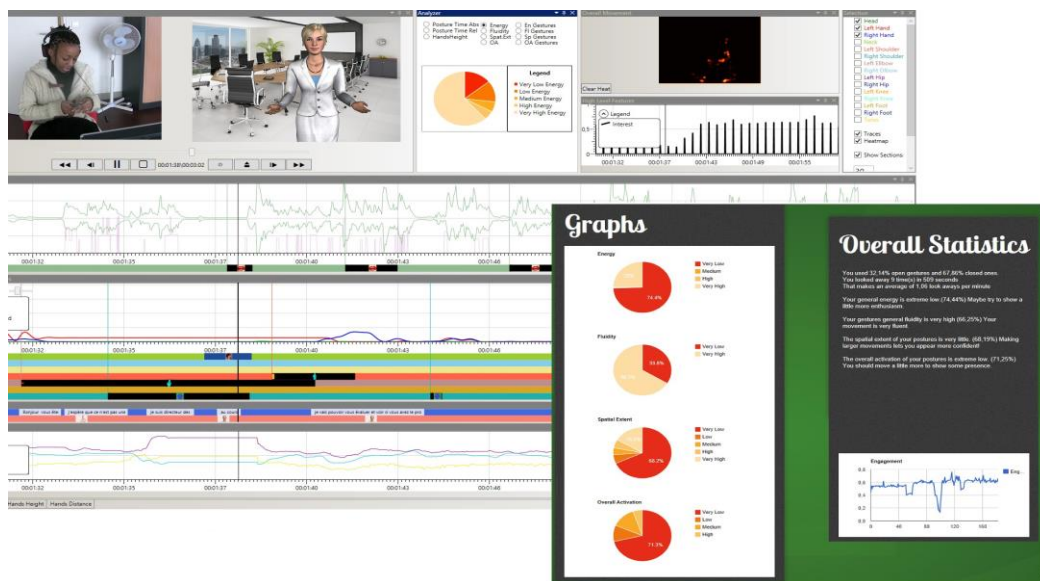


Figure 4. TARDIS scrutable OLM showing synchronised recordings of the learners interacting with the AI agents along with the interpretation of the learner's low-level social signals such as gaze patterns, gestures, voice activation in terms of higher-level judgements about the quality of those behaviours, e.g., energy in voice.

The model's assessment over these behaviours is then visualised to the learner as shown by the pie charts in Figure 4, as a way to provide the users with a concrete and immediate basis for reflecting on how they may improve their verbal and non-verbal behaviours in subsequent interviews. A time-lined view of learner actions that the system detected and

interpretation of those actions is also given. This OLM provides a detailed basis for more nuanced discussion about learners' job interview performance and specific behaviours with human practitioners than would otherwise be possible. The evaluation of this OLM showed significant improvements in key behaviours targeted, including the quality of the responses to interview questions, non-verbal behaviours such as gestures, voice modulation and eye gaze, as well as leading to learners' decreased levels of anxiety and increased levels of self-efficacy and confidence (Porayska-Pomsta & Chryssafidou, 2018). Interestingly, in line with existing OLM research, the accuracy of TARDIS' diagnoses does not seem a pre-requisite of the success of the intervention. Indeed, some inaccuracies in the model may even be desirable, if the explicit goal of the interaction with an OLM is to provoke to student to self-reflect, self-explain or argue with the system about its diagnosis. Here, the potential of OLMs as mirrors and as props for metacognitive competencies development is clearly apparent, providing a unique opportunity for EdN to study these competencies in a systematic and ecologically valid ways, for example by linking the idea of explicitly separating learners' subjective experience from the observation of the behaviour and thus, through OLM turning such observation into a more objective, almost vicarious experience.

Humanly AI and the social brain

Another important question related to cognitive fidelity of the AI models that is of relevance to EdN is that related to the definition of the 'sufficiently humanly behaviour'. In the broader context of AI, this question is central to progressing the state of the art in the field and one which has been asked since the idea of systems that both support and depend on an interaction with humans has emerged in the 1960s (e.g., Licklider, 1960; Englbart, 1962). In this context one of the more intriguing hypotheses is the uncanny valley hypothesis (Mori, 1970, 2005), which states that humanlike objects, for example some forms of robots, elicit our emotional responses, e.g., empathy, similar to those that are elicited in response to other humans. Although, the degree of the emotional responses to such objects tends to be proportionate to the degree of human likeness, beyond a certain degree of similarity and realism, such responses can suddenly become extremely negative (Misselhorn, 2009). Over decades, this hypothesis has led to substantial AI research investigating the questions of AI models' socio-emotional and behavioural credibility in human-computer interaction and of the relationship between human users' empathy towards and social affinity with technological objects that may be attributed some human qualities (e.g., Slater, 2006). AI researchers have focused on finding the necessary and sufficient human-like characteristics (their appearance as well as verbal and non-verbal behaviours) of AI agents in a variety of application contexts and with different users, including different educational applications (e.g., Moreno et al., 2001; Baylor & Kim, 2004), social interactions (e.g., Pelachaud & Andre, 2010), and special needs interventions such as autism (e.g., Porayska-Pomsta et al., 2018).

The questions of credibility relate to both the physical appearance and behaviour of the agents, as well as their seeming trustworthiness as experts in a given learning domain and as educational practitioners. For example, with respect to physical appearance of educational AI agents, Baylor investigated the impact of gender (female, male) and ethnicity (e.g., White, Black), role (e.g., expert, motivator, mentor), and realism (e.g., realistic, cartoon) of the agents on learning transfer, self-regulation and self-efficacy. Their results showed that students had greater transfer of learning when the agents were more realistic

and when they were represented non-traditionally (as Black versus White) when in the “expert” role. Many studies also suggest that when agents are perceived by the learners as less intelligent, this can lead to significantly improved self-efficacy, whereas the use of motivational messages, as employed through the motivator and mentor agent roles, can lead to enhanced learner self-regulation and self-efficacy (Baylor, 2004).

Neuroscience research is also beginning to shed light on the apparent similarities between the neural processes that occur when we engage with other humans vs. when we interact with human like AI. Here the questions tackled also relate to the sufficiently humanly behaviour that is needed to trigger our attributing human intentionality, i.e., theory of mind (ToM) to AI (Howard-Jones, 2009). For example, an fMRI study by Krach et al. (2008) suggested that visual appearance is critical in increasing such attributions, showing increased activation in the participants’ brain regions that are associated with ToM (Figure 6) the more a piece of technology appears to be human. Here the experimental conditions included a computer, a functional robot, an anthropomorphic robot, and a human – see Figure 5. While the study did not address the questions related to the uncanny valley hypothesis, namely whether and at what point increased realism may lead to deactivation of the ToM regions and occurrence of feelings of negativity towards hyper-realistic agents, it does suggest that investment in human-like qualities may be important to learners’ mental engagement with AIEd as tools for supporting learning.





Interaction Partner				
Condition	Computer Partner (CP)	Functional Robot (FR)	Anthropomorphic Robot (AR)	Human Partner (HP)
Humanlikeness	no human shape; no perceivable button pressing	no human shape; button pressing with artificial hands	humanlike shape; button pressing with humanlike hands	human shape; button pressing with human hands

Figure 5. Four conditions examined by Krach et al. (2008).

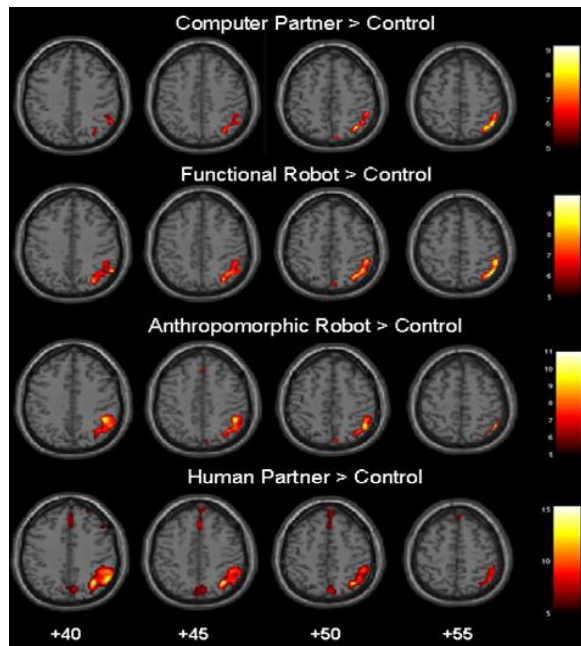


Figure 6. Regions associated with ‘theory of mind’ grow more active as the appearance of a technological opponent becomes more human-like, even when it is clearly not human (Krach et al. 2008).

However, more realistic agents may not be desirable in all learning contexts. For example, in the ECHOES project, which developed an interactive system for supporting social interaction skills acquisition by children with autism, a cartoonish agent was employed (Figure 7). Here, we placed the emphasis on creating a socially credible, but evidently non-human agent in an attempt to remove some of the social anxiety associated with autism, while at the same time exaggerating some of the features such as the agent’s eyes and emotional displays (e.g., surprise, happiness). The ECHOES agent proved to be an effective social partner to children leading to increases in their initiations of and responses to bids for social interaction during the use of the ECHOES environment (Porayska-Pomsta et al., 2018).

The same types of cartoonish agents as used in ECHOES are presently employed in another project, called unLOCKE, where we investigated the impact of a computerised neuroscience intervention on primary school children’s ability to learn counterintuitive facts in maths and science (main intervention) and on their understanding of socially challenging scenarios (active control) (e.g., Wilkinson et al., in press). In both conditions, children observe agents’ actions: in the main intervention four agents are placed in a TV-like game show settings where they have to answer questions related to counterintuitive science and maths problems. Three agents act as contestants, whereas one agent acts as the show’s host. Children first observe how the contestant-agents respond to the challenges posed to them by the host-agent, who also confirms which answer is a correct one. Following this observation phase, children can attempt some problems of their own. In the active control condition, the agents engage in social interaction with one another around key topics such as bullying, or social exclusion, before the child is asked to analyse the social scenarios they observe using targeted prompts from the system. In both ECHOES and unLOCKE, the computational modelling relates to agents’ behaviours, which must be contingent on the pedagogical goals of any given learning scenario, on the state of the world inhabited by the

agents (other agents, objects, etc.), and in the case of ECHOES – on the actions of the users on the environment. This is achieved through the application of GOF AI planning architecture, which responsible for managing the agents' immediate reactions and deliberative actions, as well their emotional displays (Dias & Paiva, 2005).



Figure 7. A child playing with the ECHOES agent through the multi-touch screen interface (Left). The agent points to a flower that it wants a child to pick and put in the basket in a bid for attention and interaction with the child (Right).

Both ECHOES and unLOCKE facilitate vicarious engagement with the respective systems, with ECHOES further allowing children to imitate the agent's actions within the environment and to engage in joint attentional activities with the agent. There is emergent evidence from EdN research supporting the value of both human-like technologies, which aligns with AI research to date, and of employing such technologies to facilitate vicarious learning such as facilitated in ECHOES and unLOCKE. Specifically, studies have shown that observing others performing actions causes the neural activation in the same cortical areas (the mirror neuron system activation) as those that occur when we are carrying out actions ourselves (Rizzolatti & Craighero, 2004). However, the mirror neuron activation seems to be restricted to human movement, suggesting that animation is most conducive to learning when it involves human movement (Tversky & Morrison 2002; Howard-Jones, 2009). What is not clear from these studies, and what might become an interesting area of study at the intersection of AIEd and EdN, is where the boundaries between credibly human and clearly non-human movement and behaviours lie, and how the different degrees of humanness can be used to support learning in different learning contexts and with different learner populations. Additionally, there is a scope for substantial research involving EdN methods which focuses on the neural activations of learners engaging in self-inspection and self-regulation with the help of OLMs.

Conclusion

We have presented two complementary approaches to learning and teaching that both employ computational methods. The first approach, building computational models of cognition, is analytical, in that it relies on developmental cognitive neuroscience theories to identify key components involved in education-relevant abilities. Modelling brings theoretical advances through clarity at the expense of simplification, and provides a platform to consider how the constraints of brain function impact cognition. The multiple components identified in the analytic approach hint at the true complexity of learning in the classroom, while the mechanistic understanding that is the goal nevertheless still requires pedagogic insights to achieve translation into classroom practices.

The second AI approach demands a fuller picture of the learner, the domain to be learned, an appropriate pedagogy for teaching the domain, and ways communication can deliver those aims. As in the first approach, the net result is a push for explicitisation and clarification of theory. We saw a range of computational tools available to support teaching and learning, with less need that the computational systems are faithful to constraints of neurocomputation, but a much greater need that systems respect the reality of real-world learning situations. AI in education, in that sense, has the potential to act as a bridge between educational neuroscience research and real-world educational practices.

Computational methods in cognitive science are one tool amongst many, a tool with strengths (rigour) and weaknesses (simplification). We need to ensure that the simplifications intrinsic in computational models do not impact on breadth of questions that are considered within educational neuroscience, in service of its ambition to utilise a mechanistic understanding of mind to achieve wider evidence-informed approaches to educational methods and policy making.

References

- Aleven, V., & Koedinger, K.R. (2013). Knowledge component approaches to learner modeling. In *Design Recommendations for Adaptive Intelligent Tutoring Systems*, volume 1 of *Learner Modeling*, pp. 165–182. US Army Research Laboratory, Orlando, Florida, R. Sottolare, A. Graesser, X. Hu, & H. Holden (eds.), 2013. ISBN 978- 0-9893923-0-3.
- Aleven, V., McLaren, B M., Sewall, J., van Velsen, M., Popescu, O., Demi, S., Ringenberg, M., & Koedinger, K R. (2016). Example-Tracing Tutors: Intelligent Tutor Development for Non-programmers. *International Journal of Artificial Intelligence in Education*, 26(1):224– 269, March 2016. ISSN 1560-4306. doi: 10.1007/s40593-015-0088-2. URL
- Azevedo, R., & Aleven, V. (eds). (2013). *International Handbook of Metacognition and Learning Technologies*, Springer International Handbooks of Education.
- Baker, R. (2010). Data mining for education, *International Encyclopedia of Education* 7(3), 112-118.
- Baker, R., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining* 1(1), 3-17.
- Basu, S., Biswas, G., & Kinnebrew, J. S. (2017). Learner modeling for adaptive scaffolding in a computational thinking-based science learning environment. *User Modeling and User-Adapted Interaction*, 27(1):5–53, March 2017. ISSN 0924-1868. doi: 10.1007/
- Baylor, A., & Kim, Y. (2004). Pedagogical Agent Design: The impact of agent realism, gender, ethnicity, and instructional role, in *Proceedings of the International Conference on Intelligent Tutoring Systems*, pp. 592-603, Springer.
- Beal, C. R. (2013). AnimalWatch: An Intelligent Tutoring System for Algebra Readiness, in *International Handbook of Metacognition and Learning Technologies*, Springer International Handbooks of Education 26, DOI 10.1007/978-1-4419-5546-3_22.
- Blackburne, L. K., Eddy, M. D., Kalra, P., Yee, D., Sinha, P., & Gabrieli, J. D. (2014). Neural correlates of letter reversal in children and adults. *PLoS one*, 9(5), e98386. doi:10.1371/journal.pone.0098386
- Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, 108(3), 624-52.
- Botvinick, M., & Plaut, D. C. (2004). Doing without schema hierarchies: A recurrent connectionist approach to normal and impaired routine sequential action. *Psychological Review*, 111, 395-429.
- Box, G.E.P. & Draper, N.R. (1986). *Empirical Model-building and Response Surface*. John Wiley & Sons, New York, NY.
- Bull, S. (1995). 'Did I say what I think I said, and do you agree with me?': Inspecting and questioning the student model, in *Proceedings of the 7th World Conference on Artificial Intelligence in Education*, 1995.
- Bull, S. & Kay, J. (2016). SMILI: a framework for interfaces to learning data in open learner models, learning analytics and related fields. *International Journal of Artificial Intelligence in Education*, 26(1):293– 331, Mar 2016. ISSN 1560-4306. doi: 10.1007/s40593-015-0090-8. URL <https://doi.org/10.1007/s40593-015-0090-8>.
- Campbell, J. I. D. (1994). Architectures for numerical cognition. *Cognition*, 53, 1-44.
- Chen L., Lambon Ralph, M. A., & Rogers, T. T. (2017). A unified model of human semantic knowledge and its disorders. *Nat Hum Behav*. 2017 Mar;1(3). pii: 0039. doi: 10.1038/s41562-016-0039. Epub 2017 Mar 1.

- Conati, C., Porayska-Pomsta, K., & Mavrikis, M. (2018). AI in Education needs interpretable machine learning: Lessons from Open Learner Modelling, *CML Workshop on Human Interpretability in Machine Learning* (WHI 2018), Stockholm, Sweden.
- Corbett, A. T., Koedinger, K. R., & Anderson, J. R. (1997). "Intelligent tutoring systems." In *Handbook of Human-Computer Interaction*, by M. G. Helander, T. K. Landauer and P. Prabhu. Amsterdam, The Netherlands: Elsevier Science.
- Cukurova, M., Luckin, R., Millán, E., & Mavrikis, M (2018). The NISPI framework: Analysing collaborative problem-solving from students' physical interactions, *Computers and Education*, vol. 116, pp. 93-109, Pergamon
- Davis, R., Shrobe, H., Szolovits, P. (1993). What is knowledge representation? *AI Magazine* 14(1), 17–33.
- Dehaene, S. (2003) The neural basis of the Weber–Fechner law: a logarithmic mental number line. *Trends in Cognitive Sciences*, 7, 145-147.
- Dehaene, S. (2005). Evolution of human cortical circuits for reading and arithmetic: The "neuronal recycling" hypothesis. In S. Dehaene, J.R. Duhamel, M. Hauser, G. Rizzolatti (Eds.), *From Monkey Brain to Human Brain*, (pp. 133-157). Cambridge, MA: MIT Press.
- Dehaene, S., & Cohen, L. (1995). Towards an anatomical and functional model of number processing. *Mathematical Cognition*, 1, 83-120.
- Dias, J., & Paiva, A. (2005). Feeling and reasoning: A computational model for emotional characters. In: *Progress in Artificial Intelligence*. pp. 127–140.
- Elman, J. L. & McRae, K. (2017). A model of event knowledge. In Gunzelmann, G., Howes, A., Tenbrink, & T., Davelaar, E. (Eds.), *Proceedings of the Thirty-Ninth Annual Meeting of the Cognitive Science Society* (pp. 337-342). Austin, TX: Cognitive Science Society.
- Elman, J.L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett K. (1996). *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.
- Engelbart, D. C. (1962). *Augmenting human intellect: a conceptual framework*. Summary Report AFOSR- 3233, Stanford Research Institute, Menlo Park, CA.
- Filippi, R., Karaminis, T., & Thomas, M. S. C. (2014). Language switching in bilingual production: Empirical data and computational modelling. *Bilingualism: Language and Cognition*, 17(2), 294-315.
- Gopnik, A., & Bonawitz, E. (2015). Bayesian models of child development. *WIREs Cogn Sci* 2015, 6:75–86. doi:10.1002/wcs.1330
- Haarmann, H., & Usher, M. (2001). Maintenance of semantic information in capacity-limited item short-term memory. *Psychonomic Bulletin & Review*, 8, 568-578.
- Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review*, 106, 491–528.
- Harm, M. W., & Seidenberg, M. S. (2004). Computing the Meanings of Words in Reading: Cooperative Division of Labor Between Visual and Phonological Processes. *Psychological Review*, 111(3), 662–720. DOI: 10.1037/0033-295X.111.3.662
- Harm, M. W., McCandliss, B. D., & Seidenberg, M. S. (2003). Modeling the successes and failures of interventions for disabled readers. *Scientific Studies of Reading*, 7, 155–182.
- Hernandez-Orallo, J., & Vold, K. (2019). *AI Extenders: The Ethical and Societal Implications of Humans Cognitively Extended by AI*, Association for the Advancement of Artificial Intelligence.

- Hoffman, P., McClelland, J. L., & Lambon Ralph, M. A. (2018). Concepts, control, and context: A connectionist account of normal and disordered semantic cognition. *Psychological Review*, *125*(3), 293-328. doi: 10.1037/rev0000094.
- Howard-Jones, P. (2009). Neuroscience, learning and technology (14-19), *BECTA Report*.
- Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can Machines Think? Interaction and Perspective Taking with Robots Investigated Via Fmri', *PLoS ONE*, *3*.7, e2597.
- Lewandowsky, S. (1993). The rewards and hazards of computer simulations. *Psychological Science*, *4*, 236-243.
- Li, N., Cohen, WW., Koedinger, KR., & Matsuda, N. (2011). A machine learning approach for automatic student model discovery, *Proceedings 4th International Conference on Educational Data Mining*.
- Licklider, J. C. (1960). Man-computer symbiosis. *IRE transactions on human factors in electronics* (1):4-11.
- Long, Y., & Aleven, V. (2013) Supporting Students' Self-Regulated Learning with an Open Learner Model in a Linear Equation Tutor. In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Proceedings of the 16th International Conference on Artificial Intelligence in Education, AIED 2013*, (pp 219-228), New York: Springer. doi: 10.1007/978-3-642-39112-5_23.
- Mabbott, A., & Bull, S. (2006) Student preferences for editing, persuading, and negotiating the open learner model. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems, ITS'06*, pp. 481-490, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3-540-35159-0, 978-3-540-35159-7. doi: 10.1007/11774303_48.
- Macfadyen LP., Dawson, S., Pardo, A., & Gasevic, D. (2014). Embracing big data in complex educational systems: The Learning analytics imperative and the policy challenge. *Research & Practice in Assessment*, *9*, 17-28.
- Mareschal, D. & Shultz, T. R. (1999). Development of children's seriation: A connectionist approach. *Connection Science*, *11*(2), 149-186
- Mareschal, D. & Thomas M. S. C. (2007) Computational modeling in developmental psychology. *IEEE Transactions on Evolutionary Computation (Special Issue on Autonomous Mental Development)*, *11*, 137-150.
- Mareschal, D., Butterworth, B. & Tolmie, A. (2013). *Educational neuroscience*. Oxford, UK: Wiley Blackwell.
- Mareschal, D., Johnson, M., Sirios, S., Spratling, M., Thomas, M. S. C., & Westermann, G. (2007). *Neuroconstructivism: How the brain constructs cognition*. Oxford: Oxford University Press.
- Martinez Maldonado, R., Kay, J., Yacef, K., & Schwendimann, B. (2014). An Interactive teachers' dashboard for monitoring groups in a multi-tabletop learning, *International Conference on Intelligent Tutoring Systems*, 482-492, Springer.
- Mavrikis, M. (2008). Data-driven modelling of students' interactions in an ILE, *Educational Data Mining*.
- McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*, 419-457.

- McCloskey, M. (1991). Networks and Theories: The Place of Connectionism in Cognitive Science. *Psychological Science*, 2(6), 387–395. <https://doi.org/10.1111/j.1467-9280.1991.tb00173.x>
- McLeod, P., Plunkett, K., & Rolls, E. T. (1998). *Introduction to connectionist modelling of cognitive processes*. New York, NY, US: Oxford University Press.
- Misselhorn, C. (2009) Empathy with Inanimate Objects and the Uncanny Valley, *Minds & Machines* (2009) 19:345–359 DOI 10.1007/s11023-009-9158-2.
- Moreno, R., Mayer, R E., Spires, H. A., Lester, J C. (2001), The Case for Social Agency in Computer-Based Teaching: Do Students Learn More Deeply When They Interact with Animated Pedagogical Agents? *Cognition and Instruction* 19(2), pp. 177-213, Lawrence-Erlbaum Associates, Inc.
- Mori, M. (1970). Bukimi no tani, *Energy* 7(4), 33–35, translated into English by K.F. MacDorman and T. Minato (2005). *Proceedings of the Humanoids-2005 workshop: Views of the Uncanny Valley*, Tsukuba, Japan.
- Mori, M. (2005). On the uncanny valley. *Proceedings of the Humanoids-2005 workshop: Views of the Uncanny Valley*, Tsukuba, Japan.
- O'Reilly, R. C., Bhattacharyya, R., Howard, M. D., & Ketz, N. (2014). Complementary learning systems. *Cognitive Science*, 38, 1229–1248. DOI: 10.1111/j.1551-6709.2011.01214.x
- Pelachaud, C., & Andre, E. (2010). Interacting with Embodied Conversational Agents, in *Speech Technology*, pp. 123-149, Springer Verlag.
- Plaut, D. C., McClelland, J. L., Seidenberg, M., & Patterson, K. E. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, 103, 56–115.
- Porayska-Pomsta, K. (2016). AI as a methodology for supporting educational praxis and teacher metacognition, *International Journal of Artificial Intelligence in Education*, Vol.26(2), 679-700.
- Porayska-Pomsta, K., & Bernardini, S. (2013). Learner Modelled Environments, in *The SAGE Handbook of Digital Technology Research* (pp. 443-458), doi: 10.4135/9781446282229.n30
- Porayska-Pomsta, K., & Chryssafidou, E. (2018), Adolescents' Self-regulation During Job Interviews Through an AI Coaching Environment, *International Conference on Artificial Intelligence in Education*, 281-285, Springer Cham.
- Porayska-Pomsta, K., & Mellish C. (2013). Modelling human tutors' feedback to inform natural language interfaces for learning, *International Journal of Human-Computer Studies*, 71(6), pp. 703-724, Academic Press.
- Porayska-Pomsta, K., & Rajendran, T. (2019). Accountability in human and artificial intelligence decision-making as the basis for diversity and educational inclusion. In *the Speculative Futures for Artificial Intelligence and Educational Inclusion*, Springer Nature – AICFE Future Schools 2030 book series.
- Porayska-Pomsta, K., Alcorn, A. M., Avramides, K., Beale, S., Bernardini, S., Foster, M-E., Frauenberger, C., Pain, H. Good, J., Guldborg, K., Kea-Bright, W., Kossyvakis, L., Lemon, O., Mademtzi, M., Menzies, R., Rajendran, G., Waller, A., Wass, S., & Smith, T. J. (2018). Blending human and artificial intelligence to support autistic children's social communication skills, *ACM Transactions on Human-Computer Interaction*, in press.
- Porayska-Pomsta, K., Rizzo, P., Damian, I., Baur, T., André, E., Sabouret, N., Jones, H., Anderson, K., & Chryssafidou, E. (2014) Who's afraid of job interviews? definitely a

- question for user modelling. In Dimitrova, Vania, Kuflik, Tsvi, Chin, David, Ricci, Francesco, Dolog, Peter, and Houben, Geert-Jan (eds.), *User Modeling, Adaptation, and Personalization*, pp. 411–422, Cham, 2014. Springer International Publishing. ISBN 978-3-319-08786-3.
- Rajalingham, R. Issa, E. B., Bashivan, P., Kar, K. et al. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *The Journal of Neuroscience*, *38*(33), 7255–7269.
- Richardson, F. M., Seghier, M. L., Leff, A. P., Thomas, M. S. C., & Price, C. J. (2011). Multiple routes from occipital to temporal cortices during reading. *Journal of Neuroscience*, *31*(22), 8239–8247. doi:10.1523/JNEUROSCI.6519-10.2011.
- Ritter, F. E., Tehrani, F., & Oury, J. D. (2018). ACT-R: A cognitive architecture for modeling cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, e1488.
- Rizzolatti, G., & Craighero, L. (2004), 'The Mirror Neuron System', *Annual Review of Neuroscience*, *27*, 169–92.
- Russell, S. J., & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach*. Prentice Hall. Second Edition.
s11257-017-9187-0. URL <https://doi.org/10.1007/s11257-017-9187-0>.
- Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*, 523–568.
- Shultz, T. R. (2003). *Computational Developmental Psychology*. Cambridge, MA: MIT Press.
- Slater, M., et al. (2006). A virtual reprise of the Stanley Milgram obedience experiments. *PLoS ONE*, *1*(1), e39. doi:10.1371/journal.pone.0000039.
- Spencer, J. P., Perone, S., & Buss, A. T. (2011). Twenty years and going strong: A dynamic systems revolution in motor and cognitive development. *Child Dev Perspect*. 2011 December ; *5*(4): 260–266. doi:10.1111/j.1750-8606.2011.00194.x.
- Spencer, J., Thomas, M. S. C., & McClelland, J. L. (2009). *Toward a new unified theory of development: Connectionism and dynamical systems theory re-considered*. Oxford: Oxford University Press.
- Storrs, K., Mehrer, J., Walther, A., & Kriegeskorte, N. (2017). Architecture matters: How well neural networks explain IT representation does not depend on depth and performance alone. Poster presented at the *Cognitive Computational Neuroscience conference*, New York, USA. (Retrieved from <https://www2.securecms.com/CCNeuro/docs-0/5928796768ed3f664d8a2560.pdf> 17 September 2019)
- Thomas, M. S. C. & McClelland, J. L. (2008). Connectionist models of cognition. In R. Sun (Ed.), *Cambridge handbook of computational cognitive modelling*. Cambridge: Cambridge University Press
- Thomas, M. S. C., Fedor, A., Davis, R., Yang, J., Alireza, H., Charman, T., Masterson, J., & Best, W. (2017). Computational modeling of interventions for developmental disorders. *Psychological Review*, 2019 Jun 6. doi: 10.1037/rev0000151. [Epub ahead of print]
- Thomas, M. S. C., Forrester, N. A., & Ronald, A. (2013). Modeling socio-economic status effects on language development. *Developmental Psychology*, *49*(12), 2325–43. DOI:/10.1037/a0032301.
- Thomas, M. S. C., Forrester, N. A., & Ronald, A. (2016). Multi-scale modeling of gene-behavior associations in an artificial neural network model of cognitive development. *Cognitive Science*, *40*(1), 51–99. DOI: 10.1111/cogs.12230

- Thomas, M. S. C., Mareschal, D., & Dumontheil, I. (2020). *Educational Neuroscience: Development Across the Lifespan*. London, UK: Psychology Press.
- Thomas, M.S.C, Ansari, D., & Knowland, V.C.P. (2019). Annual Research Review: Educational neuroscience: progress and prospects. *Journal of Child Psychology and Psychiatry*, 60(4), 477–492. doi:10.1111/jcpp.12973
- Tversky, B., and Morrison, J.B. (2002) Animation: Can It Facilitate?, *International Journal of Human-Computer Studies*, 57, 247-62.
- Ueno, T., Saito, S., Rogers, T. T., & Lambon Ralph (2011). Lichtheim 2: synthesizing aphasia and the neural basis of language in a neurocomputational model of the dual dorsal-ventral language pathways. *Neuron*, 72(2), 385-96. doi: 10.1016/j.neuron.2011.09.013.
- Westermann, G., Mareschal, D., Johnson, M. H., Sirois, S., Spratling, M. W., & Thomas, M. S. C. (2007). Neuroconstructivism. *Developmental Science*, 10(1), 75-83.
- Wilkinson, H. R., Smid, C., Morris, S., Farran, E. K., Dumontheil, I., Mayer, S., Tolmie, A., Bell, D., Porayska-Pomsta, K., Holmes, W., Mareschal, D., & Thomas, M. S. C. (in press). Domain-specific inhibitory control training to improve children’s learning of counterintuitive concepts in mathematics and science. *Journal of Cognitive Enhancement*.
- Wolff, B. (2008). *Building Intelligent Tutoring Systems*. Morgan Kaufman.
- Ziegler, S., Pedersen, M. L., Mowinckel, A. M., & Biele, G. (2016). Modelling ADHD: A review of ADHD theories through their predictions for computational models of decision-making and reinforcement learning. *Neuroscience and Biobehavioral Reviews*, 71, 633–656.
- Zorzi, M., Stoianov, I., & Umiltà, C. (2005). Computational Modeling of Numerical Cognition In: J. Campbell (Ed.), *Handbook of Mathematical Cognition* (p. 67-84). New York: Psychology Press.