Rethinking mental representation through the epistemology of modeling

Draft, 14.03.2026

Renne Pesonen, Tampere University, renne.pesonen@gmail.com orcid:0000-0001-6425-5772

Abstract:

The problems associated with classical symbolism need not entail abandoning representationalism. Indeed, while representational theories of mind have been widely challenged by embodied and action-oriented approaches, representationalism remains alive and well in empirical research on higher cognition, where the notion of conceptual representation appears indispensable. I argue that a reconciliation between representationalism and action-oriented theories of cognition is possible through a pragmatist interpretation of mental model research. Mental models are flexible knowledge structures that integrate information at multiple levels of abstraction to serve diverse epistemic and pragmatic functions. By combining this research with the epistemology of scientific modeling, I propose replacing symbolic referentialism with pragmatist inferentialism, an account of representation grounded in notions of inference and adequacy for purpose, rather than reference and truth conditions. This framework dissolves the divide between philosophical and psychological theories of concepts and aligns more closely with current empirical research on higher cognition than either classical symbolism or its radical anti-representational alternatives.

1. Introduction

The received view in the cognitive sciences is that thinking and reasoning require mental representations. Classical formulations of representationalism assert that mental representations are symbolic, forming language-like combinatorial structures with conceptual content (Fodor, 1987; Fodor & Pylyshyn, 1988). What gives mental symbols their content is their ability to refer to the things, properties, and propositions our thoughts are about. While classical symbolism has lost traction in recent decades, the notion of reference as a prerequisite for mental content remains central to many philosophers, including both advocates and critics of the representational theory of mind (e.g., Hutto & Myin, 2013). Critics often emphasize embodied and action-oriented approaches, which tend to focus on neural and sensorimotor representations (e.g., Clark, 2013) or reject the notion of representation altogether. Anti-representationalist sentiments typically stem from the close conceptual link between representation and symbolic reference. Consequently, representationalism is frequently assumed to stand or fall with the symbolic paradigm.

In practice, radical anti-representationalism has gained little traction in empirical research on higher cognition, where the notion of representation, including structured conceptual representations, remains alive and well. However, this has not forced cognitive scientists to formulate their theories in terms of mental symbols or the Language of Thought hypothesis. A growing number of cognitive scientists believe that the notion of mental models is key to understanding higher cognition. Mental models flexibly incorporate knowledge at various levels of specificity, from sensorimotor to abstract, to support perception, action, learning, reasoning, and language. In this paper, I combine resources from this research and the epistemology of scientific modeling to develop an empirically motivated pragmatist account of mental representation that is compatible with both combinatorially structured conceptual content and action-oriented theories of cognition.

My argument proceeds as follows: Traditional symbolic referentialism imposes constraints on conceptual representation that are too strict and empirically vacuous. Clinging to referentialism has led some philosophers to treat philosophical and psychological concept research as separate enterprises. In Section 2, I argue this is an error. Instead of abandoning psychology or the notion of mental representation, we should replace symbolic referentialism, derived from the philosophy of language, with pragmatist inferentialism, derived from the epistemology of scientific modeling and presented in Section 3. Section 4 argues that this shift enables us to vindicate the notion of mental representation with a richer, empirically informed theory of mental models. Mental and scientific models share many important similarities in their representational properties, suggesting that models constitute a generic representational type or style particularly well suited to theorizing about a wide range of human cognitive activities.

Mental models retain aspects of symbolic representations, namely the explicit combinatorial representation of objects and relations over mere statistical knowledge and pattern matching (see Lake et al., 2017). I do not argue against these aspects of symbolism, but against symbolic referentialism, which further requires that mental content be determined by determinate reference and truth conditions. According to pragmatic inferentialism, representation is not an intrinsic property of the model–target relation, appraised by factors such as veridicality, accuracy, or structural similarity. Instead, almost anything can be used to represent anything if the representational vehicle is adequate for its user for some particular purpose. Likewise, mental models should not be understood as accurate images or descriptions of reality. Rather, they are flexible knowledge structures that support a range of pragmatic and epistemic functions.

However, the analogy between scientific and mental models only goes so far. While scientists can stipulate what their models represent, this strategy is not available to theories of mental representation. This disanalogy is taken up in Section 5, which explains how models are acquired and selected through adaptive learning, drawing on research in model-based reinforcement learning. On this view, the function of a model is to support inferences about the current situation and thereby to represent the task and its context. Section 5 also argues that mental representations of categories are best understood as repositories of knowledge that guide model construction and model-based inference across contexts, rather than as representations of conceptual essences.

Amalgamating research on mental models with the epistemology of modeling allows us to formulate a pragmatist notion of mental representation, grounding it not in reference and truth but in inference and adequacy for purpose. It also clarifies why pragmatic context is more fundamental to mental representation than accurate reference.

2. Referentialism and its problems in the philosophy of mind

2.1. Symbolic referentialism

Both mental images and mental symbols have been repeatedly proposed as the format of mental representations. Image theory has its roots in the old empiricist idea that mental contents can be traced to perceptions that can be later mentally invoked for thinking and reasoning. Apart from the problem of how to represent abstract and formal concepts with images, the theory is plagued by problems of indeterminacy: The same image can represent many things, and things with similar appearances may belong to entirely different categories. These problems paved the way for the widespread adoption of the symbolic theory of mental representation in the philosophy of mind and the cognitive sciences.

In addition to providing an account of mental representation, classical symbolism includes a specific theory of cognitive processes: Thinking and reasoning are syntactically defined computations over quasi-linguistic symbol structures. When the brain processes these symbols, we make conceptual inferences. This explains how abstract and counterfactual reasoning is possible, and how we can think about things not present in immediate experience (Haugeland, 1991; Clark, 1997). In conjunction with computationalism, symbolism thus explains how thinking can possess both the expressive power of language and the inferential power of logic (Fodor & Pylyshyn, 1988).

However, the symbolic theory of mental representation inherits some key problems from its roots in the analytic philosophy of language—namely, what gives mental symbols their representational character and grants them the conceptual content they supposedly have? Since the defining property of symbols is reference, the default answer to these questions is some form of referential semantics, which concerns how words are paired with aspects of the world to assign meaning to linguistic expressions and determine the truth-values of sentences.

2.2. Problems with referentialism

Perhaps the most direct approach to grounding mental symbols in aspects of the world comes from causal-informational theories of content. The idea is that if the perception of a particular thing or property reliably causes the activation of a specific brain state, then that state functions as a representation of the perceived thing or property (Fodor, 1987, Ch. 4). However, pure causal theories are widely considered insufficient because purely causal notions cannot account for representational error, which is a prerequisite for referential theories of content (Fodor, 1987; Hutto & Myin, 2013).

Consequently, many theorists have turned to teleosemantics, which resembles causal theories but additionally holds that human cognition is composed of systems whose proper functions involve the production and consumption of representations. For example, perception produces representations that are consumed by decision-making processes. This allows for the definition of representational error in naturalistic terms as an instance of malfunction. However, if representational content is defined by accurate reference and determinate truth-conditions, teleosemantic theories seem not to fare much better than pure causal-informational theories (Hutto & Myin, 2013, Ch. 4). After decades of research, teleosemantics still struggles with problems of inadequacy and indeterminacy in content attribution, even in simple cases (Bergman, 2023).

According to inferentialist alternatives, mental content depends on how concepts are used in thinking and reasoning, instead of causal contact with referents. But again, the problems inherited from the philosophy of language loom large. Individuals employ all kinds of idiosyncratic beliefs in their routine inferences, and there must be a way to distinguish conceptually correct inferences from incorrect ones. Unless this issue is resolved, pure inferentialism risks rendering everyone’s conceptual systems mutually incommensurable, thereby amplifying the problem of indeterminacy to the extreme. A solution would require a clear distinction between conceptual and empirical

knowledge, but this necessitates invoking a strict analytic/synthetic distinction long abandoned by philosophers and psychologists (Fodor & LePore, 1991).

Some inferentialists resort to two-factor theories in which some referents are fixed by the causal route and some by the inferential route (e.g., Block, 1986). This gives more room for dealing with error and indeterminacy, because even if individuals have different inferential contents, they may share the same referent through perceptual contact, or *vice versa*. However, if the two factors are always perfectly aligned, one becomes redundant. If they are not, the theory fails to specify which factor determines reference in each case (Perlman, 1997). An additional account for fixing the referent thus seems to be required. Yet that is precisely what the theory is supposed to accomplish.

Presumably, both perceptual and inferential content are important for conceptual representation. Hence, perhaps the two-factor theory could be worked out satisfactorily, for example, by treating the choice between the factors as an empirical question. However, I doubt this approach because I doubt the viability of referentialist semantics for mental representation for the following reasons.

Referentialists are preoccupied with determining how the mind tracks abstract objects, such as properties and categories, with determinate conceptual identities. This is supposed to secure the propositional contents of thought and our capacity to have beliefs and desires. I doubt the viability of this approach, since it is rarely adopted in cognitive psychology. There, concepts are widely regarded as malleable, contextual, and idiosyncratic bodies of knowledge that underwrite higher cognitive competencies. Explanations in empirical cognitive science virtually never hinge on representational content in the sense intended by symbolic referentialists—as fixed, mind-independent universals.

Some have concluded that psychology is irrelevant to philosophical theories of mental representation because psychologists are not investigating concepts but something else (Machery, 2009). I believe this conclusion is misguided. Next, I argue that philosophical research on concepts should not be divorced from empirical psychology. Note that my primary concern is mental content; my arguments therefore do not necessarily count against externalist theories of linguistic meaning. At the very least, I agree with the core tenet of externalism: If anything determines referents, it is not internal mental states.

2.3. Concepts and conceptions

Edouard Machery has characterized the concept of "concept" within cognitive science as follows: "A concept of *x* is a body of knowledge about *x* that is stored in long-term memory and that is used by default in the processes underlying most, if not all, higher cognitive competences when these processes result in judgments about *x*" (Machery, 2009, 12).

Machery goes on to analyze the division of labor between philosophers and psychologists. Psychologists are concerned with the bodies of knowledge involved in higher cognitive processes, while philosophers are interested in aboutness: What makes our judgments actually be about *x*, and how we can have propositional attitudes, such as beliefs and desires, about anything in general? Cognitive psychologists generally presume that reference is either unproblematic or irrelevant to their explanatory interests. Hence, philosophers and psychologists have different aims, and by "concept" they mean different things, which should be kept apart (Machery, 2009, Ch. 2).

Gregory Rey (1983) insists that psychologists are not investigating concepts but *conceptions*. According to Rey, concepts are determined by factual descriptions of essences, or by the necessary and sufficient conditions for category membership. This was also the dominant account of concepts in psychology until the adoption of feature-matching theories in the 1970s, according to which concept representations consist of a set of characteristic, rather than defining, features. Even if defining features did exist, they are typically not used as the basis for categorization (Hampton, 2006; Murphy, 2002). For Rey, this shift marked a parting of ways between philosophers and psychologists, because psychology turned to investigating only epistemological and linguistic functions, while concepts also function as the basis for metaphysical taxonomy—by which I mean a universal, mind-independent category system that determines how each concept is individuated, irrespective of the epistemological question of how they are acquired.

Metaphysical taxonomy is important for Rey and other referentialists, such as Fodor (1998), because there must be some standard that determines the correct and incorrect ways of using and understanding concepts. Semantics concerns how conceptions mirror concepts, and metaphysical taxonomy provides a universal standard that is also supposed to explain the inter- and intrapersonal stability of conceptual meaning. This stability, in turn, is regarded as necessary for the ability of different individuals to share propositional attitudes with the same contents.

2.4. Psychology matters

While the distinction between concepts and conceptions is instructive, psychological research has demonstrated beyond doubt that human cognition does not track metaphysical essences or analytic truths. Conceptual knowledge generally relies on generic descriptions rather than definitions (Hampton, 2006), and violations of extensional logic are common in taxonomic inference (Sloman, 1998). Concepts, whether mental or linguistic, are not stable universals but malleable and often discontinuous. This holds true in individual development from childhood to adulthood, historically from one conceptual system to another (Nersessian, 2008; Carey, 2009), between novices and experts, and even across different contexts within the same individual (Barsalou, 1987). Even if a metaphysically privileged taxonomy exists, it remains an abstraction, insulated from actual minds and languages. We mortals must manage with mere conceptions, and therefore so must psychological explanations.

Some concepts arguably have clear definitions that individuals know and use, including certain theoretical and abstract concepts especially in mathematics. However, I believe it was a mistake of early analytic philosophy to take the philosophy of mathematical logic as a model for concepts in general, because what works for mathematical concepts clearly does not work everywhere. Well-defined, formal concepts should be regarded not as paradigm cases but as exceptions. Accounting even for them presumably requires attention to the linguistic and other social practices from which they originate. For example, *an even number* is a simple, well-defined concept, yet our grasp of it arises not from a rational intuition but from the complex interplay of innate cognitive abilities and cultural practices (Carey, 2009).

While the scope of referential theories—and classical symbolism in particular—may be limited to explaining propositional attitudes and not the nature of higher cognition, the division of labor between philosophy and psychology is not as clear-cut as Machery (2009) suggests. For instance, the main selling point of classical symbolism is its ability to explain the systematicity and productivity of thought (Fodor & Pylyshyn, 1988). The theory clearly involves a rough but genuine empirical hypothesis about the nature of higher cognition, and it is frequently treated as such in the literature.

Beliefs and desires are widely taken to be dispositions that manifest in cognitive and overt behavior. The symbolic referentialist strategy for explaining how we can have propositional attitudes is to attribute conceptual content to mental representations, which are consumed by cognitive processes implementing these characteristic dispositions. This strategy simply cannot work insulated from the

psychological reality of those representations. In particular, the explanatory aim cannot be fulfilled by postulating internal representations that track an abstract, mind-independent conceptual realm, because mental representations clearly track something else. Psychology matters because we need to know what they actually track and how.

Instead of relying on a metaphysical taxonomy, there are other ways to set semantic standards and explain the stability of meaning, for example, shared biology, psychology, environment, goals, cultural practices, and social norms. Because these are shared only to some extent, the resulting semantic standards are contextual and intersubjective rather than universal and objective (Tomasello, 2008), blurring the distinction between concepts and conceptions. Unfortunately, the sociocultural aspects of mental content cannot be explored in full here. Nevertheless, in the following sections, I argue that abandoning symbolic referentialism does not require rejecting the notion of representational content, since a better theory of representation is available.

3. Model-based pragmatic inferentialism

While research on mental representation has enjoyed a long and close relationship with the philosophy of language, alternative accounts can be sought from the contemporary philosophy of science. Since the early 2000s, the study of scientific representation has focused almost exclusively on models (Knuuttila, 2011; Frigg & Nguyen, 2022). This section examines model-based inferentialism, according to which the representational properties of models arise from their epistemic properties—specifically, from their inferential affordances within epistemic practices. By contrast, referentialists maintain that epistemic properties arise from more fundamental representational properties. In Section 4, this account is applied to formulate a theory of mental representation.

Models used in science are highly diverse, ranging from scale and data models to agent-based computer simulations. My discussion primarily addresses causal, mechanistic, and simulation models, as they prove relevant to mental representation. I focus on a few epistemological aspects of modeling that are important to our later discussion of mental models. By committing to inferentialism and pragmatism, I am not implying that these tenets cannot be reasonably contested in theories of scientific representation. Rather, I adopt them as premises for developing a theory of representation that is useful in the context of mental models.

3.1. Inferentialism and scientific modeling

Epistemological questions of modeling are often considered somewhat distinct from representational questions. However, I advocate an epistemological answer to the representational question, namely inferentialism (Suárez, 2004; Kuorikoski & Ylikoski, 2015; Khalifa et al., 2022), according to which a model can be taken to represent a phenomenon or system insofar as it enables its users to make predictive, explanatory, heuristic, or other inferences about the modeled target. Inferentialism highlights a particularly important epistemological property of models, namely *surrogate reasoning.* Especially when the target system is too small or large, too distant or complex, or otherwise inconvenient to investigate directly, one can gain insights into the system by implementing some relevant assumption about it in a model and studying the model system instead. This also allows one to study hypothetical systems that do not necessarily exist.

In Section 2.2, we already encountered problems with inferentialism. Specifically, it is difficult to pinpoint the exact set of inferences a representation should allow in order to count as the representation of its target. For scientific models, this requirement primarily concerns factual rather than conceptual inferences. The problem with scientific models is that they are often highly simplified and idealized surrogates of their supposed targets. From the Volterra–Lotka equations to agent-based models in the social sciences, models omit many properties of their targets and introduce features that are not actually there. They typically idealize, simplify, and distort to the extent that they become literally false descriptions of their targets.

Nevertheless, even highly idealized models can be very useful inferential tools for many epistemic and pragmatic purposes. Beyond accurate representation, scientific models often serve other functions, such as facilitating understanding, explanation, and communication. That is, accurate representation is just one possible function of models among many—not a logical prerequisite for fulfilling the rest—and excessive accuracy may even hinder other purposes (Potochnik, 2015; Kuorikoski & Ylikoski, 2015). In particular, when the goal of modeling is to isolate and study a selected property of a complex target, glossing over even important details can produce an artificial system that fulfills important explanatory or heuristic functions, even if it behaves unlike any real-world system.

Consequently, if representation is determined solely by inferential affordances, referents remain underdetermined because models almost never afford complete and exclusively sound inferences about their putative targets. However, this problem stems from an exclusive focus on model–target

relations and is arguably not critical for the pragmatist interpretations of inferentialism discussed next.

3.2. Pragmatism in scientific representation

Attempts to define the representational relation of specific models in isolation, based solely on their inferential affordances or, for example, on model–target resemblance, consistently lead to problems that can be alleviated by including the model user and purpose in the analysis (see Frigg & Nguyen, 2022). According to such *pragmatist* inferentialism, there is no point in asking whether model M is the representation of target T as such. Instead of focusing solely on the dyadic model–target relation, we must always factor in the model user, purpose, and context of use, and model evaluation should be based on an assessment of adequacy for purpose rather than (mere) accuracy or veridicality (Giere, 2010; Parker, 2020; Khalifa et al., 2022). According to this view, representation is not a property of models as such but emerges from the activity of model users.

For example, you can use a saltshaker, a pepper mill, and a mug to represent the relative locations of the Lithuanian cities Vilnius, Kaunas, and Klaipėda. It would be inappropriate to ask whether the pepper mill is *really* a representation of Kaunas. In this context, it is; in most other contexts, it is not. Representational status can be invoked simply by stipulation, through which just about anything can be used to represent anything else (Callender & Cohen, 2006; Giere, 2010). Stipulation may involve more complex explanations that guide how the model should be interpreted, and it should be understood as embedded in wider activities, such as scientific and linguistic practices (Khalifa et al., 2022).

However, mere stipulation will not suffice. According to inferentialism, representational potential is constituted by inferential affordances, and that potential comes in degrees (Suárez, 2004). In principle, even if no part of the model corresponds to anything in reality, it cannot fail to represent its target if it enables some useful inferences about the target for some purpose in a given context. In practice, however, it would be virtually impossible for a model to support correct, non-trivial inferences without capturing some important aspects of the target, such as causal structure or relevant explanatory mechanisms. Nevertheless, veridical correspondence is not a logical prerequisite for representation, and neither is accuracy, which becomes a pragmatic requirement for adequate representation. According to Kuorikoski and Ylikoski (2015, 3830): “models represent only some aspects of the modeled systems, and the kinds of inferences made using the model

determine what these aspects are and the extent to which these inferences are correct determines how accurate the representation is."

Many modeling techniques are reused and adapted to investigate phenomena very different from those for which the models were initially developed (Callender & Cohen, 2006; Knuuttila & Loettgers, 2016). This travel of modeling templates highlights the importance of pragmatic contexts that guide the interpretation and use of models, as it demonstrates how inadequate mere model–target correspondences are for explaining how models are used for scientific representation. For example, the Volterra–Lotka predator–prey model was derived from chemistry (by Lotka) and mechanics (by Volterra), and it later found its way into economics. The Ising model, developed for studying ferromagnetism, has found applications in disciplines as remote as the social sciences (Knuuttila & Loettgers, 2016). It may be that ferromagnetism and social opinion formation truly share some abstract dynamic commonality that makes the Ising model adequate for studying both phenomena. Nevertheless, because the actual models are highly idealized, their mere formal properties leave completely undetermined which specific systems or phenomena they describe.

To summarize: While models are clearly representations, they are not things that are true or false in the sense that they intrinsically refer, or fail to refer, to aspects of the world. Instead, their representational contents are rooted in inferential affordances that enable us to reason that certain states of affairs are thus and so. Given that concepts are bodies of knowledge that allow us to form judgments about certain states of affairs (see Section 2.3), models clearly satisfy this criterion and hence function as conceptual representations (see also Nersessian, 2008). According to pragmatic inferentialism, representation also depends on the context of use, and representational use requires interpretation grounded in other cognitive functions that support the inferential uses of models.

Moreover, models often have a combinatorial structure, in that they have parts that can be added, removed, and modified. However, this property differs from propositional compositionality, whereby the contents of sentences are inherited from logical combinations of their atomic parts. With models, the primary unit of representation is the model itself, which delineates the identity of its components. This is particularly evident in scientific models that can be interpreted in terms of abstract relational categories (see Kokkonen, 2017). While models can have other models as parts, atomic model components, in isolation, have no interpretation until put together and put to use.

3.3. Modeling as extended cognition

Finally, to begin forming a link between scientific modeling and mental models, I draw on theoretical views that conceptualize models as external cognitive tools (Nersessian, 2008; Knuuttila, 2011; Kuorikoski & Ylikoski, 2015). For example, Knuuttila (2011) emphasizes that model systems are artifacts that support various cognitive functions as concrete, manipulable objects. I make no commitment to the ontological claim that models *are* artifacts. However, this perspective helps to explain certain key cognitive functions of modeling that are important for model-based mental representation in general.

Another set of theoretical ideas I draw on holds that thinking operates through the mental simulation of mental models. Such simulations support everyday planning and decision-making (Gilbert & Wilson, 2007; Baumeister et al., 2016), as well as theoretical and hypothetical reasoning through thought experiments (Nersessian, 2017). John Epstein (2008) captures this idea neatly: “[W]hen you close your eyes and imagine an epidemic spreading, or any other social dynamic, you are running some model or other. It is just an implicit model that you haven't written down.”

When you write down your intuitive mental model, you do not faithfully replicate what is inside your head. You create a new kind of object, the external model, that enables inferences through manipulating, adapting, and evaluating it (Nersessian, 2008). Such models can range from *ad hoc* sketches to more elaborate constructs, such as complex computer simulations.

The way the model system is concretely constructed partly determines its inferential affordances and constraints. Using various representational means, it is possible to represent, manipulate, and communicate ideas and information that can be difficult to verbalize. For example, sketches allow one to use spatial and visual cues to represent the structural and dynamic aspects of the target. Of course, inferential affordances and constraints are also partly determined by the cognitive skills of the model user. Hence, model-based inferences involve the interplay of external and internal representations and processes, making modeling a prime example of extended cognition.

Modelers gain new insights into a model’s behavior by experimenting with it, modifying it, exploring its properties under different parameterizations, and so on. These interactions enable users to develop an intuitive understanding of the model’s properties and qualitative behavior (Kuorikoski & Ylikoski, 2015). This learning is based on the acquisition of new mental models and cognitive skills that allow users to reason about the external model by mentally simulating the acquired internal model (Nersessian, 2008). Through these interactions, the external and internal models

become coupled: The external model may initially function as an explication of an intuitive mental model, but through interactional engagement, it becomes a source of new mental models and inferential skills that supports increasing understanding of its properties and behavior.

Three points arise from this discussion. First, the representational model–target relation is virtually irrelevant for the aspect of modeling examined in this subsection, which strictly concerns procedural learning through model–user interaction. For such learning, it does not matter whether the model is intended to represent anything at all.

Second, under pragmatist inferentialism, it appears unproblematic to say that the acquired mental model represents the external model: It is acquired through practices involving reasoning about the external system, and it enables the user to make inferences about the external model, even when the latter is unavailable.

Third, the mental model inherits some of the representational properties of the external model, to the extent that it enables some of the inferences supported by the external model. That is, users become capable of reasoning not only about the external model but also with it.

Next, we turn to research on mental models. After a brief overview, we discuss mental models in light of the pragmatic model-based inferentialism examined in this section.

4. Mental models, mental simulation, and model-based inferentialism

4.1. A brief overview of mental models and mental simulation

There is no single theory of mental models. The notion appears in many research programs, some of which are only remotely connected. The original idea in modern cognitive science comes from Kenneth Craik, who conjectured that organisms carry in their heads small-scale models of external reality and their own possible actions, enabling them to investigate and prepare for future situations before they arise (Craik, 1943, 61). In the early 1980s, the notion was applied in research on sentential deductive reasoning (Johnson-Laird, 1983) and on domain-specific skills and knowledge, abstraction, and analogical reasoning (Gentner & Stevens, 1983).

No consensus exists on whether mental models reside in the long-term memory or working memory, or on the exact relationship between them and other cognitive representations, such as

schemata. Nevertheless, a common idea is that schemata in the long-term memory encode generic predictive knowledge about highly familiar situations. Mental models, by contrast, are specific knowledge structures that combine multiple schemata (along with other background knowledge) in a way that enables agents to mentally represent and simulate specific events, including hypothetical scenarios (Jones et al., 2011). Mental models are thought to be incomplete and inaccurate depictions of things and events, and they evolve according to individuals' goals, skills, and experience. They embody intuitive understanding that "trades accuracy and veridicality for speed, generality, and ability to make predictions that are good enough" (Battaglia et al., 2013, 18328).

Many theorists emphasize that mental models involve perceptual knowledge. For example, according to Johnson-Laird (2008, 47), "each mental model represents a possibility in as an iconic way as possible." They also incorporate spatial and temporal information, as well as causal, mechanical, and procedural knowledge into mental simulations (Nersessian, 2017). Mental imagery presumably activates background knowledge in ways analogous to perception. That is, imagery serves both as construction material for models and as a probe for contextually relevant knowledge, along with affective responses that guide goal selection (Barsalou, 1999; Gilbert & Wilson, 2007).

In short, mental models are best understood as flexible knowledge structures that opportunistically integrate information across multiple levels of specificity. Their function is not to provide accurate or veridical representations; rather, their primary role is to guide situated action by supporting inference, decision-making, planning, and judgment. Furthermore, they enable surrogate reasoning through the simulation of hypothetical or absent events (Koechlin, 2014; Nersessian, 2017). Thus, while scientific and mental models differ in many respects, their properties as representations appear highly similar under pragmatist inferentialism. In particular, the representational properties of mental models are rooted in their inferential affordances, given the specific goals of the agent.

4.2. Combining model-based inferentialism with mental models

Section 4.1 argued that scientific and mental models share important qualitative similarities in their representational properties. The analogy between the two is not perfect, however. How it breaks down depends on the specific kind of inferentialism applied to mental models, which cannot be identical to that applied to scientific modeling. I address this issue in Section 4.3. First, let us examine in more detail how mental models can be combined with model-based pragmatist inferentialism and its implications for the nature of mental representation.

Section 3.3 argued that through experimenting and interacting with concrete model systems, users develop cognitive skills that facilitate the better intuitive understanding of the model. Apart from procedural learning, this understanding is based on the acquisition of a mental model of the external model. Here, we use this schema to understand mental representation in general. First, we drop the representational model–target relation from consideration and focus solely on user–model interactions. Then, we allow the environment to take the role of the external model in any pragmatic context. The implication is that goal-directed interactions with the environment generally function in exactly the same way as interactions with external model systems. That is, activity in any recurring context leads to the development of a mental model of that context, which underwrites agents' understanding of it, along with the associated inferential and pragmatic skills.

The last claim above is essentially the cornerstone of the mental models research. So why to derive it from theories of scientific modeling, and why go through all the discussion of the model–target relation if we ultimately dismiss it and instead turn our focus to agent–environment interactions and mental models? The reason is that we use the *scientific model–target* representation relation to conceptualize the *mental model–environment* representation. The implications are numerous:

1. The claim that mental models represent the environment is too loose. They represent contextually relevant aspects of the environment for the agent for specific tasks or purposes.
2. Mental models are the primary unit of mental representation. They guide the interpretation of the general gist of a context and the specific things and events embedded within it.
3. The representational properties of mental models are rooted in their inferential affordances, which can be more or less complete and accurate depending on the model and agent's cognitive skills. Accuracy is a contextual measure of adequacy, not a logical precondition for representation.
4. Pragmatic utility and adequacy are more important than veridicality or accuracy. By focusing on what is contextually relevant, mental models streamline information, thereby facilitating explanation, understanding, and qualitative prediction.
5. The brain does not encode particular things and events in exactly the same way across different contexts or individuals. Instead, models can vary widely depending on the agent's experiences, goals, knowledge, and skills within specific contexts.
6. Models and their components can be reused for conceptualizing completely new scenarios as well as reconceptualizing familiar ones. In particular, familiar abstract schemata can be used through analogical transfer for making inferences about poorly understood contexts.

7. When no adequate model is available, one can be generated *ad hoc* for the purpose at hand. Any background knowledge and analogies can be used in its construction.
8. Models can be further refined and adapted. This may involve fine-tuning for better empirical fit or more radical conceptual restructuring as the function of developing domain knowledge.

Most items on this list should sound familiar to those acquainted with embodied and action-oriented approaches to cognition. However, we have arrived at them from an entirely different perspective—namely, the epistemology of modeling, which deals with explicit, combinatorially structured conceptual representations. This mutual coherence suggests that the account developed here can conceptually accommodate both action-oriented and higher conceptual cognition.

4.3. On the analogy between scientific modeling and mental representation

I close this section by detailing the version of inferentialism that I believe best applies to mental representation. A good starting point is the deflationary approach developed by Suárez (2004), according to which there is no need for a substantial theory of representation, because the concept does not do any explanatory work in the inferential use of models in science. All the general truths pertaining to their representational properties are captured by two conditions that are necessary, though not jointly sufficient, for model M to represent target T: (1) the representational force of M points towards T, and (2) M allows competent and informed agents to draw specific inferences regarding T. Condition (2) is the surrogate reasoning condition, which allows that M can be anything, as long as it enables agents to draw any kind of informative inferences about T (Suárez, 2004, 773). It should be clear by now that this condition applies also to mental models.

Suárez's approach has been criticized for leaving both the nature of the required inferences and the notion of representational force ambiguous (Frigg & Nguyen, 2022, Ch. 3). Nevertheless, Suárez insists that these notions admit of no informative and general analysis with respect to representation, although something more substantial can be said about them in each specific instance where models are used (Suárez & Solé, 2006). If inferentialism is to be informative for the philosophy of mind, something more substantial must be said about conditions (1) and (2).

Regarding the nature of the inferences in condition (2), the considerations presented in Section 2 license us to leave that question to empirical psychology: It is an empirical question what kinds of inferences mental models allow and which kinds of knowledge and mechanisms they employ.

However, some rough ideas have already been introduced in Section 4.1, and more will follow in Section 5.

Most cognitive psychologists (see Murphy, 2002), along with traditional referentialists such as Fodor (1998), consider concepts to be categories. Hence, in Section 5.3, I situate category research within the inferentialist framework to preempt counterarguments that my main argument ultimately fails to illuminate conceptual representation and instead addresses something else.

Regarding the "representational force" in condition (1), Suárez (2004) suggests that in scientific modeling it typically amounts to stipulation, which is thought to link external models to their targets via underlying mental representations (Callender & Cohen, 2006) or linguistic practice (Khalifa et al., 2022). To avoid circularity, we need an account of mental representation that does not hinge on such conditions.

Accordingly, we replace stipulation with another type of intentional activity: goal-directed, interactional engagements with the environment that underlie the acquisition of mental models. This idea further coheres with action-oriented theories of cognition in implying that representational content depends on intentional activity. In Section 5.1, I develop this idea in more detail by drawing on model-based reinforcement learning, which also serves as a useful theoretical background for the discussion of category representation in Section 5.3.

5. Procedural learning, mental models, and category representation

In Section 4, we identified a disanalogy between scientific and mental models: On pain of circularity, we cannot appeal to underlying mental representations or stipulation to explain the "representational force" of mental models. This notion refers to some contextual factor that leads a competent user to consider a specific target and thereby determines what the user's inferences concern (Suárez, 2004).

While linguistic practices surely can serve as sources for mental models (see Section 3.3), our aim is to show that mental models can independently function as conceptual representations in virtue of their role in inferences. Hence, accounting for "representational force" amounts to explaining what guides agents to employ specific models in drawing inferences about particular targets, without invoking language or other conceptual representations.

My explanation appeals to the interactional engagements that underlie the acquisition of mental models. To make this proposal more informative, I discuss model-based reinforcement learning in the context of human learning and decision-making. I emphasize that this discussion is intended to illustrate how the proposal can be developed in more concrete terms, not to provide a theoretical account of how it must be implemented.

5.1. Mental models and reinforcement learning

Reinforcement learning is one of the main areas of research in computer and cognitive science that links adaptive learning, decision-making, and probabilistic models. It was initially inspired by behaviorist learning theories and later found application in explaining the role of dopamine neurons in reward prediction in the brain (Sutton & Barto, 2018). Today, it provides a general framework for complex adaptive learning through interaction with the environment.

Reinforcement learning involves learning a policy for choosing actions that maximize cumulative reward over time. At each time step, the agent observes the current state of the environment, selects an action, receives a reward signal, and transitions to a new state. By repeating this cycle of interaction, the agent gradually improves its behavior, estimating through trial and error which actions are more valuable in which states, without requiring an explicit model of how the environment works. Because agents aim to maximize long-term reward, action selection is not based solely on choosing the most immediately rewarding action in each state. Through exploration, agents can learn that even punishing actions may be valuable if they lead to higher long-term returns than immediately rewarding options. Powerful learning algorithms have been developed that enable agents to acquire highly complex action policies through entirely model-free learning, relying only on estimates of the long-term value of each action in each state.

Psychologically, reinforcement learning provides a sound theoretical framework for habit learning and more complex skill acquisition. However, humans—and even rats—do not choose actions solely based on learned *stimulus→action* values, but also in a goal-oriented way, based on learned *action→outcome* associations (Drummond & Niv, 2020). This becomes evident when goals change or when we need to plan our actions. In such cases, behavior cannot be based solely on stimuli and action values, but it instead requires a model of the environment that supports goal-directed inference.

For example, if you are suddenly assigned a new task at work, the stimulus environment may remain the same while your goals change, rendering your previous stimulus → action associations maladaptive. In principle, one solution to adjust to this change is to keep learning, updating value estimates as you go, and trying to keep pace with the changing environment. In practice, this approach is unsuitable for quickly changing contexts because cached values adjust slowly, while rewards can change abruptly when goals change. Worse yet, new learning overwrites the old, and if the previous context reoccurs, you have lost valuable, hard-earned skills you now need to relearn.

Apparently, the human brain solves the problem by keeping track of different pragmatic contexts in addition to mere stimulus–action–outcome contingencies, which can vary widely across contexts. During learning, we automatically seek a more abstract, higher-level context under which to group individual actions, even when tasks are simple and there is no clear behavioral benefit to doing so (Collins et al., 2014). The long-term benefit lies in the ability to associate each context with its own set of *action → outcome* associations that provide a predictive model of the task environment. When pragmatic contexts change, the agent can switch models and adjust action policies through model-free learning (Domenech & Koechlin, 2015).

As there are potentially infinite contexts in the open environment, the brain now faces a new problem: How should a model be assigned for each external context, and, if it generates incorrect predictions, should the model be refined through further learning or replaced entirely, because the agent is actually facing a novel situation? This is an inductive problem that requires the brain to keep track of different contexts and align them with the available background knowledge.

The brain's solution appears to be to select a single model that is consistent with contextual cues and then use it to continuously generate inferences about what is likely to happen, while monitoring the reliability of these inferences. Because these models are not strictly determined by stimuli, multiple models consistent with the external situation can be retrieved for a given context. If the active model consistently errs, it is deemed unreliable and replaced with an alternative that more accurately predicts action outcomes. If no ready model is available in the long-term memory, a new one can be created using mixtures of previously learned behavioral strategies. If that fails, the agent must learn an entirely new strategy and its associated predictive model through trial and error (Koechlin, 2014; Domenech & Koechlin, 2015).

According to Etienne Koechlin (2014), the human prefrontal cortex consists of three integrated but functionally and evolutionarily discernible layers. The paralimbic prefrontal cortex is responsible for selecting active action policies and monitoring their reliability. This neural system is reactive and depends entirely on factual feedback received during activity. The next layer, the lateral prefrontal cortex, is sensitive not only to action outcomes but also to contextual cues that signal changes in external contingencies. This neural structure harbors contextual models and enables action control based on proactive inferences before acting. Lastly, the evolutionarily newer frontopolar regions support planning as well as hypothetical and counterfactual reasoning by allowing us to mentally create and manipulate alternative models.

5.2. Interim summary

The discussion above implies that prediction and goal-directed action are supported by mental models acquired through interactive engagement with the environment. The sole function of these models is to enable inferences about relevant external contingencies, allowing agents to achieve their goals across diverse pragmatic contexts. The "representational force" that assigns mental models to each context (and, derivatively, a context to each model) is an inductive process that relies on these same inferences, guided by context cues, background knowledge, and active goals.

The ability to mentally create and simulate such models extends their representational capacities to "off-line" surrogate reasoning, which many philosophers (e.g., Haugeland, 1991) consider to be the hallmark of mental representation.

To be sure, the research cited in Section 5.1 primarily concerns how agents learn action policies instead of the rich mental simulations discussed in Section 4.1. Nevertheless, research amalgamating such simulations with reinforcement learning is underway (Lake et al., 2017; Hamrick, 2019).

One may worry that the present account cannot sustain the distinction between correct and incorrect inferences, which was invoked several times in Section 3. However, the theory discussed in Section 5.1 is essentially a version of the predictive coding hypothesis (Clark, 2013), according to which the brain continually generates expectations at multiple levels of abstraction and matches them against observations. An error signal is generated when observations diverge from predictions at the sensorimotor level or at some higher conceptual levels not tied to specific stimuli. In such cases, the model has generated an incorrect inference. Crucially, error detection relies on feedback, but it is

not directly sensitive to the mismatch between models and reality but between observations and inferences generated from the model.

There remains one disanalogy between scientific and mental models that we need to address: The latter are not employed by agents in the same way as the former. Rather, mental models are part of the agent's cognitive makeup. Here, we can invoke the process/content distinction, or the distinction between representation-producing and representation-using cognitive mechanisms. As we saw in Section 5.1, learning and action selection can proceed entirely model-free, even in the presence of mental models. Models are produced and used by distinct mechanisms that support these processes. Thus, they are not strictly constitutive to agency but provide content for inferential processes. I find it philosophically innocent to say that the agent utilizes mental models for inferences (indeed, sometimes explicitly through their mental manipulation), even if the actual user of these representations is not strictly the agent, but rather executive functions within the agent. Often this utilization is simply automatic and tacit for the agent.

Next, we devote Section 5.3 to discuss mental categories. The conclusion that emerges, unsurprisingly, is that category representation does not primarily serve logical inference between propositions, but rather causal and explanatory inference in the guidance of practical activity.

5.3. Category representations and causal models

Soon after the widespread adoption of feature-matching models in the 1970s, the knowledge view of concepts emerged in the 1980s (Murphy, 2002). It became apparent that category representations are not mere clusters of correlated features. Rather, category knowledge also explains how and why features hang together, such as how wings are causally related to flying.

According to the knowledge view, concepts are structured like mini-theories that support predictive and explanatory inferences. For example, event categories help us predict and explain what is happening and why in a given situation. Knowledge associated with relatively abstract concepts, such as *heroism* or *intoxication*, may explain why, for instance, someone jumps into a swimming pool fully clothed (Murphy & Medin, 1985, 295).

Since the notion of causal explanation is central in the knowledge view of concepts, it makes sense to investigate it using philosophical theories of explanation (Lombrozo, 2006). In both cognitive science and philosophy, theories of causal knowledge and inference commonly employ causal

networks to represent causal structure (Holyak & Cheng, 2011). Contrastive counterfactual explanations, in particular (see Kuorikoski & Ylikoski, 2015), can be straightforwardly represented as Bayesian networks. Thus, it is well motivated to interpret the “mini-theories” underlying category representation in terms of causal models.

Indeed, theories of category representation as causal models have proven valuable for explaining several categorization phenomena, including analogy (Holyoak et al., 2010), causal status and coherence effects, and psychological essentialism (Rehder & Kim, 2010), which refers to the tendency to assume that many categories—particularly natural kinds and artifacts—possess an unobservable causal core that makes their members the way they are.

Within causal models, essences can be represented as hidden causes that generate a category’s observable features, and categorization can be understood as explanatory inference from observations to the presence of a category instance. Hidden causes resemble symbols in that they can serve as placeholders for unknown properties, but their contents depend not on reference, but on their role in inferences within causal models.

In Section 5.1, model selection was likewise attributed to explanatory inference from observable cues to the presence of a familiar context. The reinforcement learning literature rarely intersects with research on concepts, generally defining models as whatever stored representations an agent can use to cope with its environment. However, category representations are natural candidates for the building blocks of model construction, and category recognition then provides a plausible mechanism for model selection.

In summary, mental categories presumably serve as inductive engines that encode mixtures of experience-based perceptual, causal, and procedural knowledge (Mandler, 2004), which can be recruited for selecting, creating, and simulating particular mental models (Barsalou, 1999). This does not imply that categories logically precede the mental models. It is entirely plausible that contextual models develop first, with category induction serving to keep track of similarities and differences across changing contexts, thereby enabling generalization and flexible knowledge transfer through networks of mental models. For example, individuals can flexibly construct goal-oriented, *ad hoc* categories, such as *things to take from one's home during a fire* (Barsalou, 1983). This ability may prove useful in the unlikely event that one’s house actually catches fire, but almost certainly not because it captures the metaphysical essence of such categories.

6. Conclusions

While classical symbolic referentialism is a problematic theory of mental representation, this is not a reason to abandon the notion of mental representation, as a better alternative is available. In particular, I have promoted model-based pragmatic inferentialism in its stead, which replaces the notions of reference and truth with those of inference and adequacy for purpose. This move appears to resolve many of the problems of referentialism, since it licenses us to trade universal truth-conditions for a more flexible notion of contextual success conditions. This alternative, however, originates from theories of scientific representation, and it is not obvious that it is psychologically plausible.

In Section 4, I argued that mental and scientific models share many important similarities in their representational properties. Although they differ in many respects, these similarities suggest that models constitute a generic representational type or style particularly well suited for theorizing about a wide range of human cognitive activities, from situated action to scientific reasoning. Accordingly, it seems plausible to draw on theories of scientific modeling to illuminate mental representation, much as classical symbolicists drew on theories of formal languages. The theory of mental representation advanced in this paper is not only better substantiated by empirical psychology than symbolic referentialism, but also philosophically informative, with implications too numerous to reiterate here (see Section 4.2).

The mentioned implications, while pertaining to higher conceptual cognition, are broadly consistent with claims made by action-oriented approaches to cognition. Accordingly, the present account of representation accommodates both action-oriented theories and combinatorially structured, conceptual representations. For example, as discussed in Section 5, predictive models are already involved in low-level, situated action control. Yet the same representations also support planning and hypothetical reasoning and can be reused for the conceptualization of entirely novel tasks and contexts. This is roughly analogous to how abstract model templates in science are repurposed to conceptualize new scenarios beyond those for which they were originally developed: The abstract gist, or relational structure, is transferred from a familiar task to another that is less well understood, and adapted for a better empirical fit and specific purposes.

The account aligns with at least some externalist theories of meaning, as the broader philosophical lesson is that, if anything determines reference, it is not internal mental states. As Georges Rey (2009) puts it, “in fixing their beliefs, people make use of whatever they think works, and what works varies.” Moreover, my argument does not necessarily conflict with teleosemantic theories, provided they are willing to let go of the idea of determinate referential content (e.g., Bergman, 2023). Like teleosemanticists, I insist that human cognition includes systems whose proper function is to represent aspects of the environment, particularly its causal structure. In Section 5.3, category representation was discussed in this light. To recap the conclusion reached there: The function of category induction is not to track metaphysical essences, but to capture pragmatically relevant regularities, enabling agents to make sense of situations through the construction of mental models.

References

Barsalou, L. W. (1983). Ad hoc categories. *Memory & Cognition, 11*(3), 211–227.

Barsalou, L. W. (1987). The instability of graded structure: Implications for the nature of concepts. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization* (pp. 101–140). Cambridge University Press.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*(4), 577–660. https://doi.org/10.1017/s0140525x99002149

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *PNAS, 110*(45), 18327–18332. https://doi.org/10.1073/pnas.1306572110

Baumeister, R. F., Vohs, K. D., & Oettingen, G. (2016). Pragmatic Prospection: How and Why People Think About the Future. *Review of General Psychology, 20*(1), 3–16. https://doi.org/10.1037/gpr0000060

Bergman, K. (2023). Should the teleosemanticist be afraid of semantic indeterminacy? *Mind & Language, 38*(1), 296–314. https://doi.org/10.1111/mila.12395

Block, N. (1986). Advertisement for a Semantics for Psychology. *Midwest Studies in Philosophy, 10*, 615-678.

Carey, S. (2009). *The Origin of Concepts*. Oxford University Press.

Callender, C., & Cohen, J. (2006). There Is No Special Problem About Scientific Representation. *Theoria, 21*(55), 67–85.

Clark, A. (1997). *Being There: Putting Brain, Body, and World Together Again*. The MIT Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(6), 181–204. https://doi.org/10.1017/S0140525X12000477

Collins, A. G. E., Cavanagh, J. F., & Frank, M. J. (2014). Human EEG Uncovers Latent Generalizable Rule Structure during Learning. *The Journal of Neuroscience, 34*(13), 4677–4685. https://doi.org/10.1523/JNEUROSCI.3900-13.2014

Craik, K. J. W. (1943). *The Nature of Explanation*. Cambridge University Press.

Domenech, P., & Koechlin, E. (2015). Executive control and decision-making in the prefrontal cortex. *Current Opinion in Behavioral Sciences, 1*, 101–106. https://doi.org/10.1016/j.cobeha.2014.10.007

Drummond, N., & Niv, Y. (2020). Model-based decision making and model-free learning. *Current Biology, 30*(15), R860–R865. https://doi.org/10.1016/j.cub.2020.06.051

Epstein, J. M. (2008). Why Model? *Journal of Artificial Societies and Social Simulation, 11*(4):12. https://www.jasss.org/11/4/12.html

Fodor, J. A. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind.* The MIT Press.

Fodor, J. A. (1998). *Concepts: Where Cognitive Science Went Wrong.* Oxford University Press.

Fodor, J. A., & LePore, E. (1991). Why Meaning (Probably) Isn't Conceptual Role. *Mind & Language, 6*(4), 328–343. https://doi.org/10.1111/j.1468-0017.1991.tb00260.x

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and Cognitive Architecture: A Critical Analysis. *Cognition, 28*(1–2), 3–71.

Frigg, R., & Nguyen, J. (2022). *Scientific Representation*. Cambridge University Press. https://doi.org/10.1017/9781009003575

Gentner, D. & Stevens, A. L. (Eds.) (1983). *Mental Models.* Lawrence Erlbaum Associates.

Giere, R. N. (2010). An agent-based conception of models and scientific representation. *Synthese, 172*: 269. https://doi.org/10.1007/s11229-009-9506-z

Gilbert, D. T., & Wilson, T. D. (2007). Prospection: Experiencing the Future. *Science, 317*(8543), 1351–1354. https://doi.org/10.1126/science.114416

Haugeland, J. (1991). Representational genera. In W. Ramsey, S. P. Stich, & D. E. Rumelhart (Eds.), *Philosophy and connectionist theory* (pp. 61–89). Lawrence Erlbaum Associates.

Hampton, J. A. (2006). Concepts as Prototypes. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation: Advances in the Research and Theory, vol. 46* (pp. 79–113). Academic Press.

Hamrick, J. B. (2019). Analogues of mental simulation and imagination in deep learning. *Current Opinion in Behavioral Sciences, 29*, 8–16. https://doi.org/10.1016/j.cobeha.2018.12.011

Holyoak, K. J., & Cheng, P. W. (2011). Causal Learning and Inference as a Rational Process: The New Synthesis. *Annual Review of Psychology, 62*, 135–163. https://doi.org/10.1146/annurev.psych.121208.131634

Holyoak, K. J., Lee, H. S., & Lu, H. (2010). Analogical and Category-Based Inference: A Theoretical Integration With Bayesian Causal Models. *Journal of Experimental Psychology, 139*(4), 702–727.

Hutto, D. D., & Myin, E. (2018). *Radicalizing Enactivism: Basic Minds without Content*. The MIT Press.

Johnson-Laird, P. N. (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness.* Harvard University Press.

Johnson-Laird, P. N. (2008). *How We Reason.* Oxford University Press.

Jones, N. A., Ross, H., Lynam, T., Perez, P., & Leitch, A. (2011). Mental Models: An Interdisciplinary Synthesis of Theory and Methods. *Ecology and Society, 16*(1): 46. https://doi.org/10.5751/ES-03802-160146

Khalifa, K., Millson, J., & Risjord, M. (2022). Scientific Representation: An Inferentialist-Expressivist Manifesto. *Philosophical Topics, 50*(1), 263–291. https://doi.org/10.5840/philtopics202250112

Knuuttila, T. (2011). Modelling and representing: An artefactual approach to model-based representation. *Studies in History and Philosophy of Science, 42*(2), 262–271. https://doi.org/10.1016/j.shpsa.2010.11.034

Knuuttila, T. & Loettgers, A. (2016). Model templates within and between disciplines: from magnets to gases – and socio-economic systems. *European Journal for Philosophy of Science, 6*(3), 377–400. https://doi.org/10.1007/s13194-016-0145-1

Koechlin, E. (2014). An evolutionary computational theory of prefrontal executive function in decision-making. *Philosophical Transaction of the Royal Society B, 389:* 20130474. https://doi.org/10.1098/rstb.2013.0474

Kokkonen, T. (2017). Models as Relational Categories. *Science & Education, 26*, 777–798. https://doi.org/10.1007/s11191-017-9928-9

Kuorikoski, J. & Ylikoski, P. (2015). External representations and scientific understanding. *Synthese*, *192*(12), 3817–3837. https://doi.org/10.1007/s11229-014-0591-2

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences, 40*: e253. https://doi.org/10.1017/S0140525X16001837

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences, 10*(10), 464–470. https://doi.org/10.1016/j.tics.2006.08.004

Mandler, J. M. (2004). *The Foundations of Mind: Origins of conceptual thought*. Oxford University Press.

Nersessian, N. J. (2008). *Creating Scientific Concepts.* The MIT Press.

Nersessian, N. J. (2017). Cognitive Science, Mental Modeling, and Thought Experiments. In M. T. Stuart, Y. Fehige, & J. R. Brown (Eds.), *The Routledge Companion to Thought Experiments* (pp. 390–326). Routledge.

Machery, E. (2009). *Doing without Concepts.* Oxford University Press.

Murphy, G. L. (2002). *The Big Book of Concepts.* The MIT Press.

Murphy, G. L., & Medin, D. L. (1985). The Role of Theories in Conceptual Coherence. *Psychological Review, 92*(3), 289–316.

Parker, W. S. (2020). Model Evaluation: An Adequacy-for-Purpose View. *Philosophy of Science, 87*(3), 457–477. https://doi.org/10.1086/708691

Perlman, M. (1997). The Trouble with Two-Factor Conceptual Role Theories. *Minds and Machines, 7*(4), 495–513.

Potochnik, A. (2015). The diverse aims of science. *Studies in History and Philosophy of Science Part A, 53,* 71–80. https://doi.org/10.1016/j.shpsa.2015.05.008

Rehder, B., & Kim, S. (2010). Causal Status and Coherence in Causal-Based Categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(5), 1171–1206.

Rey, G. (1983). Concepts and stereotypes. *Cognition, 15*(1–3), 237–262.

Rey, G. (2009). Review of Doing Without Concepts. *Notre Dame Philosophical Reviews.* https://ndpr.nd.edu/reviews/doing-without-concepts/

Sloman, S. A. (1998). Categorical Inference Is Not a Tree: The Myth of Inheritance Hierarchies. *Cognitive Psychology, 35*(1), 1–33.

Suárez, M. (2004). An Inferential Conception of Scientific Representation. *Philosophy of Science, 71*(5), 767–779. https://doi.org/10.1086/421415

Suárez, M. & Solé, A. (2006). On the Analogy between Cognitive Representation and Truth. *Theoria, 21*(55), 39–48. https://doi.org/10.1387/theoria.552

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction (2nd ed).* The MIT Press.

Tomasello, M. (2008). *Origins of Human Communication*. The MIT Press.