Preprint *Word count: 8970*

Rethinking mental representation through the epistemology of modeling

Renne Pesonen, Tampere University, Finland, renne.pesonen@gmail.com

Abstract:
Problems associated with classical symbolism need not entail the abandonment of representationalism, as some advocates of action-oriented theories of cognition have concluded. This paper argues that a better theory of mental representation can be achieved by applying the epistemology of scientific modeling to research on mental models. The resulting framework integrates action-oriented approaches to cognition with explicit, combinatorially structured representations. It is empirically far better substantiated than symbolic referentialism and aligns more closely with cognitive psychological research than radical enactivist alternatives.

Keywords: Mental representation, Mental models, Epistemology of modeling, Enactivism, Referentialism

1. Introduction

The received view in cognitive science 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 to cognition that eschew the notions of symbolic content and combinatorially structured conceptual representations.

While I mostly side with the critics, I think that the notion of representation, including combinatorially structured conceptual representation, is practically indispensable to research on higher cognition. In this paper, I show that there is an empirically motivated account of mental representation that is compatible with both structured propositional representations and action-oriented theories of cognition. To achieve this, I combine resources from the epistemology of

scientific modeling and research on mental models. This allows us to formulate the notion of representation in pragmatist terms, grounding it not in reference and truth conditions, but in inference and adequacy for purpose. It also enables us to clearly argue why pragmatic context is more fundamental for mental representation than accurate reference.

Many equate the notion of representation with the property of referring, or being about something, and symbols are representational vehicles defined by that very property. Therefore, the symbolic paradigm and representationalism are often seen as flourishing and falling together. While enactivists might endorse the notion of sensorimotor representations, there is disagreement, even within the movement (e.g., Clark, 1997), about whether it is possible to reduce conceptual cognition to sensorimotor activity. In practice, radical anti-representationalism has not gained significant traction in empirical research on higher cognition, where representationalism remains alive and well.

However, representationalism in theoretical practice 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 can flexibly incorporate knowledge at various levels of specificity, spanning from sensorimotor to abstract knowledge, in order to support perception, action, learning, and reasoning. Research on mental models is also the arena where we see a rehabilitation of symbolism in the current era of deep learning—at least insofar as symbolism can be equated with the explicit combinatorial representation of objects, identity, and relations over mere statistical knowledge and pattern matching (see Lake et al., 2017). I do not argue against symbolism as such but against symbolic referentialism, which requires that mental content is determined in terms of reference, truth, and accuracy.

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 many philosophers to treat philosophical and psychological concept research as separate enterprises. That, I argue, is an error (Section 2). 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 (Section 3). This enables us to vindicate the notion of mental representation with a richer and empirically informed theory of mental models (Section 4). Beyond scientific cognition, I examine

mental models in the context of adaptive learning, drawing in particular on research in model-based reinforcement learning, and connect this perspective to category representations (Section 5).

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, basically anything can be used to represent anything, as long as 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 situated action as well as a range of pragmatic and epistemic purposes, including predictive and explanatory inference, understanding, abstraction, and communication.

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 medium 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.

Apart from 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 our brain processes these symbols, we make conceptual inferences. This enables abstract and counterfactual reasoning, as well as other thoughts about things and ideas not present in immediate experience (Haugeland, 1991; Clark, 1997). Symbolism, in conjunction with computationalism, 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 analytic philosophy of language. Namely, what gives mental symbols their representational character and grants them the conceptual content they supposedly possess? Since the defining property of symbols is reference, the default answer to these questions is some form of

referentialism. Referential semantics 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 (provided this correlation is a law-like fact; see Fodor, 1987, Ch. 4). However, pure causal theories are widely considered insufficient because they cannot account for representational error, a prerequisite for referential theories of content.

Therefore, many advocate teleosemantics, which resembles causal theories but additionally holds that human cognition consists of systems whose proper functions involve the production and consumption of representations. Systems responsible for perception and those responsible for decision-making serve as examples. This allows for the definition of representational error in naturalistic terms as an instance of malfunction. However, if representational content is defined by reference, truth, or accuracy, 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, conceptual 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 those that are not. 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.

It is plausible that both perceptual and inferential content are important for mental content. 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 because I doubt the viability of referentialist semantics for understanding mental representation**.**

Referentialists are preoccupied with determining how the mind tracks abstract objects, such as properties and categories, with distinct conceptual identities. This is supposed to secure the propositional contents of thought and our capacity to have beliefs, desires, and so on. However, I doubt the viability of this approach, as it is rarely adopted in cognitive psychology, where concepts are widely regarded as malleable, contextual, and idiosyncratic bodies of knowledge that underwrite higher cognitive competencies. Few, if any, explanations in empirical cognitive science hinge on representational content in the sense intended by symbolic referentialists—as fixed, mind-independent universals.

Some have concluded that, therefore, psychological research is irrelevant to philosophical theories of mental representation, because psychologists are not investigating concepts but something else. I think this conclusion is misguided. Next, I argue that philosophical research on concepts should not be divorced from empirical psychology. Note that I am primarily concerned with mental content, and my arguments do not necessarily count against externalist theories of linguistic meaning. I obviously 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 factual descriptions of essences or 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, Ch. 2). For Rey, this shift marked the 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.

Metaphysical taxonomy is important for Rey and other referentialists because of the problem of misrepresentation: there must be some standard that determines the correct and incorrect ways to use and understand 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 correct concept use.

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.

While the scope of referential theories—and classical symbolism in particular—may be limited to accounting for propositional attitudes, rather than taking a stand on cognitive psychological explanations, the division of labor between philosophers and psychologists 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 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—and the processes that consume them—clearly track something else. Psychology matters because we need to know what they actually track, and how.

I believe concepts are rooted in shared conceptions, which blurs the distinction between the two. Instead of relying on metaphysical taxonomy, there are other ways to set semantic standards and explain the stability of meaning; for example, shared biology, psychology, environment, cultural practices, and social norms. Because these are shared only to some extent, the semantic standards are contextual and intersubjective rather than universal and objective. In the following two sections, I argue that abandoning symbolic referentialism does not require rejecting the notion of representational content, because a better theory of representation is available.

3. Model-based pragmatic inferentialism

While the 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. This is particularly attractive for advocates of mental models. While there is no general theory of model-based representation, the study of scientific representation has almost exclusively focused on models (Knuuttila, 2011; Frigg & Nguyen, 2022). Moreover, theories of explanation and understanding, staples of the philosophy of science, prove surprisingly relevant to the psychology of concept representation. But first, let’s take a closer look at the epistemology of scientific modeling.

3.1. Inferentialism and scientific modeling

Models used in science are highly diverse, including scale models, data models, causal models, computer simulations, agent-based models, and so on. Some models are concrete objects, some are computer programs, and some are, perhaps, abstract objects or descriptions. Here, by "models," I primarily refer to causal, mechanistic, and simulation models, for reasons that will become apparent later. I highlight only a few common features of modeling that are relevant to our later discussion of mental models, but not necessarily to all scientific models. I omit questions concerning, for example, the ontology of models and instead focus on selected epistemological aspects.

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). According to inferentialism, 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 big, too distant or complex, or otherwise inconvenient to investigate directly, one can gain insights of the system by implementing some relevant assumption about it in the model and studying the model system instead. This strategy 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 often 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 serve other functions, such as understanding, explanation, and communication. Although the purpose of some models is to accurately represent their targets, that is just one function among others, not a logical prerequisite

for fulfilling the rest. Indeed, excessive accuracy may hinder other purposes, such as facilitating understanding, communication, and explanation (Potochnik, 2015; Kuorikoski & Ylikoski, 2015).

Typically, the purpose of modeling is to isolate and study a selected property of the target phenomenon. Glossing over even important details can produce an artificial system that behaves unlike any real-world system, yet still fulfills important explanatory, exploratory, or heuristic functions. However, if representation is determined solely by inferential affordances, referents remain underdetermined, since models almost never afford complete and exclusively sound inferences about their putative targets. This issue, however, is not critical for pragmatist interpretations of inferentialism, discussed next.

3.2. Pragmatism in scientific representation

The problems of fixing a referent based on mere inferential affordances lead us to *pragmatist* inferentialism, according to which there is no point of asking is model M the representation of target T as such. Model evaluation should be based on the assessment of adequacy for purpose instead of (mere) accuracy or veridicality (Parker, 2020). All attempts to define the representational relation in isolation—for example, by resorting to model–target resemblance—lead to problems that are alleviated by including the model user and purpose in the analysis (see Frigg & Nguyen, 2022). Thus, instead of focusing on the dyadic model–target relation, we must always factor in the model user, the purpose it is used for, and the context of its use (Giere, 2010; Parker, 2020).

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. The representational status is 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 vignettes or stories that guide how the model should be interpreted. Nevertheless, it is the use context that determines reference.

Therefore, representation is cheap in principle, while in practice useful representation may not be. No matter how hard you investigate the pepper mill, you will learn nothing new about Kaunas.

According to inferentialism, representational potential is constituted by inferential affordances, and that potential comes in degrees. According to Kuorikoski & 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."

Hence, in principle, even if no part of the model corresponds to anything in reality, it cannot fail to represent its target if it enables useful inferences about the target for some purpose in a given context. In practice, it would be surprising (if not impossible) if the model supported correct non-trivial inferences without capturing some important aspects of the target, such as causal structure or relevant explanatory mechanisms. Nevertheless, accuracy, however measured, is not a logical necessity for correct representation, but a pragmatic requirement for adequate representation.

It is common for scientific models to take on a life of their own and travel across research questions and disciplines (Callender & Cohen, 2006; Knuuttila & Loettgers, 2016). That is, many modeling templates and techniques are reused and adapted to investigate phenomena very different from those the models were initially developed for. For example, the Volterra–Lotka predator–prey model was derived from chemistry (by Lotka) and mechanics (by Volterra) and later found its way into economics. The Ising model, developed for studying ferromagnetism, has found applications even in disciplines as remote as the social sciences (Knuuttila & Loettgers, 2016).

The travel of modeling templates highlights the importance of pragmatic contexts that guide the interpretation and use of models, as it demonstrates how hopelessly inadequate mere model–target correspondences are for explaining how models are used for scientific representation. It may be that, for example, ferromagnetism and social opinion formation truly share some very abstract dynamic commonality, which makes versions of the Ising model adequate for studying both phenomena (see Knuuttila & Loettgers, 2016). Nevertheless, because the actual models are highly idealized, their mere formal properties leave it completely undetermined not only which specific system they may refer to, but also what kind of 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. Representation depends on the context of use, and representational use requires interpretation, grounded in other cognitive functions that support the communicative and inferential uses of models.
It should also be emphasized that models are compositionally structured propositional representations: they represent certain states of affairs as being thus and so, by explicitly representing parts that can be added, removed, modified, or exchanged. Nevertheless, 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). 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 simulation supports everyday planning and decision-making (Gilbert & Wilson, 2007; Baumeister et al., 2016) and also underpins theoretical and hypothetical reasoning by way of thought experiments (Nersessian, 2017). In the words of Josh Epstein (2008): “[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’s 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 structural and dynamical 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.

Models do not only function as a means to explicate implicit mental models. They are used to learn non-trivial and often surprising consequences of modeling assumptions. Modelers gain new insights

into a model's behavior by experimenting with it, modifying it, and exploring its properties under different parameterizations This process leads model users to learn the intuitive feel about model's properties and its qualitative behavior (Kuorikoski & Ylikoski, 2015). Through these interactions, internal and external models become coupled. The external model supports the development of new mental models and cognitive skills, which facilitate the understanding of the external model by enabling the ability to carry out intuitive mental simulations about it based on the acquired mental model (Nersessian, 2008).

Note that the representational model–target aspect is virtually irrelevant for this aspect of modeling that concerns procedural learning through model–user interaction. It should be noted, however, that scientific modeling also involves a great deal of background knowledge about the modeled target that is relevant for devising, using, and interpreting the model. For example, Nersessian (2008) discusses how Maxwell used a series of sketches and diagrams, based on analogies to vortices, idle wheels, and gears, as models of the properties of electromagnetism. The continuous adaptation and manipulation of these models led to new conceptual insight on electromagnetism and the mathematical expression of electromagnetic forces. This process was supported and constrained also by existing knowledge about electromagnetism. In general, such background knowledge can, and must, be used to manipulate scientific models and constrain their construction and adaptation.

At this point, a referentialist may remind us that a theory of reference is needed to explain how these different kinds of knowledge hang together. For example, there must be some account that explains by virtue of what did Maxwell's various models refer to the same thing as the contemporary knowledge about electromagnetism. For now, it suffices to say that it does not matter if theories, empirical data, and heuristic models refer to anything in particular, as long as they can be used to inform and constrain each other.

4. Mental models and mental representation

4.1. From the epistemology of modeling to mental representation

Above, I explained that through experimenting and interacting with the model, the user develops cognitive skills that manifest as the understanding *of* the model. According to inferentialism, understanding is a capacity to make (correct) inferences, and the acquired know-how enables modelers to understand the modeled target better *with* the model. Apart from procedural learning, this understanding is based on the acquisition of a mental model of the external model.

The above provides a schema for understanding mental representation in general. We tell almost exactly the same story, but drop the representational model–target relation from consideration. Then, we allow the environment to take the role of the external model in any pragmatic context. That is, goal-directed interactions with the environment in general function exactly the same as interactions with external model systems: activity in any recurring context leads to the development of a mental model of that context, which—along with the inferential and pragmatic skills associated with it—underwrite the agent's understanding *of* the context *with* the acquired mental model.

The above claim is basically the cornerstone of the mental models approach and, therefore, not a very remarkable theoretical achievement in itself. So why go through all the discussion about the model–target relation if we ultimately drop it and focus solely on agent–environment interactions? First, the model–target discussion can provide a general analysis of any dyadic representation that relies on external representations (including language). Second, we use what we have learned about scientific model–target representation relations to conceptualize the mental model–environment representation. The latter is our focus below, and 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 and veridicality are measures of success, not logical preconditions for representation.
4. Utility and adequacy are more important than veridicality or accuracy. Mental models streamline and compress information, enabling explanation, understanding, and qualitative prediction. An excessive number of variables can hamper these functions, making models overly complex.
5. Models can vary widely across contexts, tasks, and individuals, and their construction reflects the experiences, goals, knowledge, and skills of the agent. Hence, there is no natural representation in the sense that the brain encodes particular things and events in exactly the same way across different contexts and individuals.
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, they can be generated *ad hoc* for the purpose at hand. Apart from analogical transfer, any background knowledge can be used.
8. Mental models can be further refined and adapted. This may involve fine tuning for better empirical fit. Sometimes it involves more radical conceptual restructuring as the function of developing cognitive skills and domain knowledge.

Some work remains to be done. First, in Section 3, part of the representational account of models was attributed to stipulation and interpretation. The act of stipulation is often thought to link an external representation to its target via underlying mental representations. To avoid circularity, we also need an account of mental representation that does not rely on stipulation. To that end, we replace stipulation with the interactional engagements that underlie the acquisition of mental models of learning contexts. Second, our task is to connect the model-based account with the cognitive psychology of concepts. Since contextual mental models are the primary unit of conceptual representation, we begin there and address psychological concept research in Section 5.

4.2. 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 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 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 subject to change as a function of persons' goals, skills, experience, and background knowledge. 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 also 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.” Mental models also exploit spatial and temporal information and mental animation to incorporate causal and procedural knowledge into mental simulations (Nersessian, 2018). Mental imagery presumably activates background knowledge in a way similar to perception. That is, imagery serves both as construction material for models and as a probe for contextually relevant causal and procedural knowledge, as well as affective responses that guide goal selection (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 and enable mental simulations for causal and mechanistic reasoning, planning and decision-making, analogical reasoning, and, apparently, also category processing (see Barsalou, 1999).

4.3. Mental models and reinforcement learning

To give a bit more specific glance at the relationship between interaction and contextual mental models, I briefly discuss reinforcement learning (RL). RL is one of the main areas of research in computer and cognitive science that links adaptive learning, decision-making, and probabilistic (mental) models. Let’s first examine plain model-free RL, and then its model-based extension.

4.3.1. Model-free reinforcement learning

Model-free RL 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, however, it serves as a general framework for modeling and implementing complex adaptive learning through interaction with the environment.

A reinforcement learning problem consists of an agent and its environment. The task of the agent is to learn 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, and then receives both a reward signal and a new state of the environment. 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.

Reward is not an external object or outcome; it is an internal signal that reinforces (or suppresses) specific actions in similar future situations. Because RL agents aim to maximize their 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 reward than immediately rewarding options.

Powerful learning algorithms have been developed that allow this simple scheme to acquire very complex action policies. The key benefit of model-free RL is its computational efficiency, as action selection in each successive state can rely solely on cached action values. Complex policies can also be structured through hierarchical combinations of simpler ones, enabling control across multiple levels of abstraction. Psychologically, reinforcement learning provides a sound theoretical framework for habit learning and more complex sensorimotor 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 we switch goals or plan our actions. Outcome-oriented planning requires a model of the environment. There is also a more fundamental reason, discussed below, why contextual mental models are essential for adaptive behavior in complex, open-ended environments.

4.3.2. Model-based reinforcement learning

What happens if actions cease to yield the expected outcomes because something has changed in the environment or within the agent? For example, food may no longer be rewarding due to satiety, or certain actions might become blocked. Many things in the environment can change, and the agent may find itself in new situations, some of which appear familiar but are not. These are challenges the human cognitive system must constantly resolve in variable and open everyday environments.

One solution is to keep learning, updating value estimates as you go, and trying to keep pace with the changing environment. However, this approach is unsuitable for quickly changing contexts because cached values adjust slowly, while rewards can change abruptly when goals change. For example, if you are suddenly assigned a new task at work, the stimulus environment may stay mostly the same while the pragmatic context does not, rendering previously learned actions maladaptive. 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 singular 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* rules that provide a predictive model of the task environment. Because these models are not strictly determined by external stimuli, multiple task sets can be associated with the same external context, and existing task sets can be brought to new contexts that are not perceived as identical to familiar ones. Each task is also associated with its own set of actions, which are adjusted through model-free RL (Domenech & Koechlin, 2015).

By keeping track of different contexts, the brain now faces a new problem: there are potentially infinite contexts in the open, action outcomes are often stochastic, and mental models are incomplete. So, if things go not as expected, how does one decide whether they have an inaccurate or completely wrong interpretation of the situation? That is, how to decide whether to continue learning to gain a better grasp of the situation, or to switch to another context and task set?

The brain appears to manage this problem by selecting a single active context model that continuously generates predictions about what is likely to happen. If the model consistently errs, it is deemed unreliable and replaced with an alternative that more accurately predicts action outcomes. If no viable model is available in long-term memory, a new one can be created using mixtures of previously learned context models and behavioral strategies. If that does not help, the agent must learn an entirely new one through trial-and-error (Domenech & Koechlin, 2015). Humans, of course, also rely heavily on social learning for guidance.

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 task sets 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 hypothetical and counterfactual reasoning by allowing us to mentally create and manipulate alternative models. This makes it possible to plan and evaluate alternative behavioral strategies.

4.4. Section conclusion

Prediction and goal-directed action selection are based on the mental simulation of contextual models, which serve as representations of tasks and task environments acquired through interactive learning. When used for planning and hypothetical reasoning, these models are best characterized as representational systems for surrogate reasoning.

For many philosophers (e.g., Haugeland, 1991), planning and hypothetical reasoning are considered hallmarks of mental representation. This is because mental representations are supposed to explain how we coordinate behavior with environmental features when those features are not reliably present: we rely on internal representations instead of direct signal from the environment. However, in theories such as Koechlin's (2014), predictive mental models are already in place at the first step of the processing hierarchy. Later stages primarily contribute additional inferential capacities but do not fundamentally alter the basic representational schemes, except for the addition of context representations at the second stage, which emerge prior to planning and hypothetical reasoning.

The function of mental models is not to represent the environment truthfully and accurately, but to capture some of its contextually relevant aspects to guide action and decision-making. Accordingly, their representational adequacy does not strictly depend on model–target correspondences, but on their inferential affordances for their users in various pragmatic contexts. The ability to mentally create, manipulate, and simulate them extends these capacities to decoupled "off-line" surrogate reasoning. In conclusion, model-based pragmatic inferentialism provides a sound account of mental representation.

Finally, one might ask whether I have been talking about concept use rather than mental representation. While inferentialism blurs the line between the two, I have not directly addressed category representation, despite its central role in both psychological and philosophical research on concepts. Before the final conclusions, I will briefly examine the link between psychological category representation and mental models.

5. 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; for instance, how wings are related to flying.

According to the knowledge view, concepts are structured like mini-theories that support predictive and explanatory inferences. These inferences often extend beyond concrete entities and their observable features. For example, event categories help us predict and explain what is happening and why in a given situation. Likewise, 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 explanation is central in knowledge view of concepts, it makes sense to investigate it using philosophical theories of explanation (Lombrozo, 2006). Hence, there is a point of contact between the psychology of concepts and philosophy of science. Moreover, in both fields, theories of causal knowledge and inference commonly employ causal-network models to represent causal structure (see 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.

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). Essentialism 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.

In the variant of reinforcement-learning theories discussed in Section 4.3.2, context model selection is likewise akin to explanatory inference from context cues to the presence of a familiar context. However, the reinforcement learning literature rarely intersects with concept research, generally defining models as whatever stored representations the agent can use to cope with its environment. While I believe such liberalism is also recommendable for understanding human cognition, category representations are the obvious candidates for the building blocks of mental models.

More precisely, mental categories presumably serve as inductive engines that encode mixtures of perceptual, causal, and procedural knowledge, which can be recruited for selecting and creating particular mental models (see Barsalou, 1999). This does not imply that categories logically precede

the mental models. It is entirely plausible that, in an individual's development, contextual models emerge 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 alternative theories are 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 for that purpose. 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 4.1).

This account also naturally accommodates both action-oriented theories of cognition and combinatorially structured, explicit representations. Given the psychological interpretation of model-based reinforcement learning discussed in Section 5.3, 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 analogous to how abstract model templates in scientific modeling 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 goals.

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 (2010) 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, including its causal structure. In Section 5, 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 identify pragmatically relevant regularities in order to guide behavior and make sense of situations through the construction of contextual 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.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *PNAS, 110*(45), 18327–18332.

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.

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., & 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.

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

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

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.

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

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