Models in the Brain

Naturalizing Human Intentionality

Dan Ryder

University of British Columbia, Okanagan

Brief description of the book

The issues surrounding mental content that were intensely debated in the 1980's and 90's remained unresolved, and no new approaches have been forthcoming since. Now, intense interest in consciousness, perceptual content, concepts, and mental representation has once again brought the issue of intentionality to the fore. *Models in the Brain* aims to present a new naturalistic theory of intentionality that addresses these problems freshly with a much-needed injection of neuroscience, opening up the subject to empirical investigation in a new way.

The central idea is that the cerebral cortex is a model building machine, where regularities in the world serve as templates for the models it builds. First it is shown how this idea can be naturalized, and how the representational contents of our internal models depend upon the evolutionarily endowed design principles of our model building machine. Current neuroscience suggests a powerful form that these design principles may take, allowing our brains to uncover deep structures of the world hidden behind surface sensory stimulation, the individuals, kinds, and properties that form the objects of human perception and thought. It is then shown how this account solves various problems that arose for previous attempts at naturalizing intentionality, and also how it supports rather than undermines folk psychology. As in the parable of the blind men and the elephant, the seemingly unrelated pieces of earlier theories (information, causation, isomorphism, success, and teleology) emerge as different aspects of the evolved model-building mechanism that explains the intentional features of our kind of mind.

- A new and detailed theory for how to naturalize intentionality (a "psychosemantics") following a decade of near silence on this issue.
- Opens up the problem of intentionality to scientific investigation by linking the theory to a mechanism, allowing formulation of specific, testable hypotheses about specific representational contents.
- Aims to uncover the real nature of what folk psychological concepts refer to (beliefs, desires, thoughts, concepts, perceivings-that...) rather than to analyze current concepts, such as that of representation, as employed by scientists.
- Offers an account of folk psychological representations (contrast the frog's infamous "fly detector") and places human intentional attitudes into an evolutionary context.
- Addresses classic problems that have arisen for the naturalization project, including indeterminacy, misrepresentation, the disjunction problem, empty representation, representation of indi-

viduals, the link between reference and conceptual role, compositionality, intentional causation, swampman, self-knowledge, objectivity, and the representation of mathematical entities.

- Informed by current work in neuroscience on cortical plasticity, the neurophysiology of sensory cortex, dendritic integration, sparse coding, predictive coding, and the default-mode network.
- Develops a (literal) mental map or model approach, while also accommodating advantages of language-of-thought and embodied approaches.

Outline of Models in the Brain

Chapter 1: An empirical approach to the mind

There are good reasons for adopting a purely empirical approach in an inquiry into the nature of intentionality. First, the concepts of mind, consciousness, intentionality, and rationality are plausibly natural kind concepts. Second, constraints on a theory of mental content (or "psychosemantics") must ultimately be derived empirically in order to avoid global anti-realism.

There may in fact be many different kinds of minds, and different kinds of intentionality; we must start at home. There are two kinds of empirical strategies to take: 1) uncover the theory of representation implicit in the cognitive sciences, while ignoring folk psychology, or 2) accept the approximate truth of folk psychology, and rely on cognitive science to reveal the real natures of folk psychological states. The second strategy is underexplored and worth taking seriously. Conclusions must be tentative in this area due to the relatively fledgling science, but the overall theoretical strategy is robust.

The "models in the brain" schema, to be filled in over the course of the book, bears a striking resemblance to an old idea going all the way back to Aristotle.

Chapter 2: The problem of intentionality

Intentionality is mental aboutness. It is central to a number of related phenomena pertaining to perception or thought, including denotation, reference, correctness, accuracy, satisfaction, and truth. The basic problems facing a naturalistic account of intentionality are legion: the disjunction problem and its relatives, the distality or chain problem, the problem of optimal conditions, empty representations, Swampman, Frege's problem, holism, compositionality, normativity, and explaining behaviour. Many of these problems arise from trying to achieve a reasonable match with folk psychology. Despite the failures of earlier theories, including those based on causation, information, conceptual-cum-causal role, isomorphism, success, plus teleological versions of these, it is too early to give up on this goal in favour of non-reductive or deflationist approaches.

Chapter 3: From teleo-isomorphism to models in the brain

Many kinds of representation are best understood teleologically. In particular, a *model* has the function of being structurally similar to what it represents (teleo-isomorphism). To represent, a teleo-isomorphism must also be subjected to appropriate uses, for example filling in missing information, guiding action, or exploring possibilities.

The representational content of a model is determined by the design principles of the modeling process that produces it, as well as the template from which it is produced. Human mental representation is to be found primarily in the cerebral cortex. If the cortex is a model building machine, the appropriate research-guiding question is: what are its design principles?

Chapter 4: SINBAD: The basics

Given the fledgling state of the relevant neuroscience, any conclusions about the design principles of cortex are necessarily tentative. The SINBAD (Set of INteracting BAckpropagating Dendrites) theory of cortical learning is one intriguing proposal, and it may be used to illustrate how one can go about building a theory of human intentionality in a way that dovetails with folk psychology.

SINBAD cortical networks are like simple associative networks, except that "associations" between nodes can be mathematically (including logically) complex, involving multiple variables. Relations among nodes are built up through learning by means of a simple but powerful dendritic matching rule. This process results in the formation of a network that mirrors regularities among significant variables in the environment, namely variables that are predictively rich. A SINBAD network becomes dynamically isomorphic to regularities involving these "deep" variables, including natural kinds, individuals, and causally significant properties. In different modes of operation, a SINBAD cortical network plays the roles of filling in missing information, guiding action, or offline exploration.

While SINBAD is empiricist in spirit, it assumes there will also be innate predispositions laid overtop of the basic mechanism.

Chapter 5: SINBAD: The nitty-gritties

The SINBAD idea is implementable by means of a variety of learning rules and physiological mechanisms. It is supported by recent work on cortical plasticity, the neurophysiology of sensory cortex (especially primary visual cortex), dendritic integration, sparse coding, predictive coding, thalamic function, and the default-mode network. The SINBAD idea links up with other neural, connectionist, and mathematical modeling approaches as well (e.g. Becker & Hinton, 1992; Phillips & Singer, 1997).

(This chapter is intended for those wanting more detail on the SINBAD theory and the neuroscientific evidence for it. While evidentially important for the SINBAD version of the "model representation" hypothesis about human intentionality, it may be safely skipped without jeopardizing an overall understanding of either this hypothesis or the SINBAD-specific version of it.)

Chapter 6: Neurosemantics

The special characteristics of the SINBAD architecture that make it useful depend specifically on its interacting with "deep", predictively rich variables: sources of mutual information. If the cortex is a SINBAD network, then, it must have been *selected for* developing that kind of isomorphism, which is therefore its biological function, its basic design principle. This supports an overall teleo-isomorphism, one of the requirements for model representation.

For any particular cell that occupies its tiny part of the model, there are physical/historical facts that determine what the template was for that cell, and therefore what particular source(s) of mutual information it represents, i.e. which source(s) of mutual information the cell's activity is supposed to correspond to in the context of the overall isomorphism.

Chapter 7: Solving the classic referential problems

The foregoing points to solutions for some of the classic problems described in Chapter 2, those dealing with the reference (broadly construed) or denotation of atomic mental representations. These include misrepresentation, referential determinacy, the distality problem (i.e. SINBAD representations do not typically represent proximal stimuli), empty representation, and the link between reference and inferential role among others. In addition, the source of mutual information category can plausibly accommodate a large range of atomic referential contents that fit with our folk psychological attributions, for both perception and thought. Even very difficult cases, like colour representations in perception and thoughts of mathematical entities, have plausible SINBAD-based explanations.

Chapter 8: Representational use and the occurrent attitudes

Teleo-isomorphism is insufficient for model representation; an isomorphic structure must also be subjected to appropriate *uses*. In the SINBAD example, these uses correspond to the modes of operation that cortical networks may enter into (Chapters 4): the roles of filling in missing information, guiding action, and offline exploration. These three modes of operation implement, or maybe constitute, the occurrent attitudes: judgement, occurrent desire, and supposition respectively. (There are also perceptuo-motor equivalents: perception, motor intention, and imagination.) The type of use in question is embodied "non-representational use", where the items used need not be represented, and it may be given either a causal-role functionalist reading, or a teleological reading. Overall, the teleo-isomorphism/mode of operation distinction corresponds to content/attitude.

Chapter 9: Propositional content, inference, and psychological explanation

The reductive account of the occurrent attitude types described in Chapter 8 requires a complementary account of attitude contents: the contents of model representations that are subject to current use. Simple attitude contents fall out relatively easily, but representational composition requires the basic SINBAD mechanism to be supplemented with the fundamental neural process of *synchrony* (see Werning, 2003; Werning & Maye, 2007). The representation of possibilities and their inferential consequences occurs by means of the "exploratory mode" of operation, which implements or constitutes the attitude of supposition.

The teleo-isomorphism of the standing model structure (realized in neural connections) gives the content of the *non*-occurrent attitudes, most importantly implicit belief. The interaction between occurrent and non-occurrent attitudes yields a complete account of inference in SINBAD networks. By extending Dretske's "structuring cause" strategy (Dretske, 1988) and capitalizing on the characteristics of isomorphism, it can be seen how these attitude contents genuinely explain behaviour.

Chapter 10: Problems for teleosemantics?

A common concern with teleosemantic theories is that they will inevitably target usefulness rather than truth. In this vein, Peacocke worries about the representation of inaccessible places and times, Godfrey-Smith doubts that teleosemantics can show how truth is the *fuel* for success, Akins finds the achievement of objectivity to be mysterious in the face of our idosyncratic subjective senses, and Pietroski doubts that the theory concerns content at all. In each case, a SINBAD-based teleosemantics suggests a natural response.

The infamous swampman objection is often perceived to be a major stumbling block for teleosemantic accounts, but the global externalism entailed by a model-building approach allays any real concerns here. Any residual swampman worries can be accommodated by the isomorphism-based aspect of the explanation of behaviour described in the previous chapter, and an explanation for how our intentionality seems to us "from the inside."

Chapter 11: Compare, contrast, and connect

It must be borne in mind that the SINBAD-based account is only one possible version of the model-building hypothesis concerning human intentionality. The approach might be fruitfully integrated with other theories of brain function, from low-level ART-based theories (Grossberg, 2013) to high-level forward-modeling (Grush, 2004) or Bayesian causal models (Gopnik et al., 2004).

Important connections or contrasts can also be made with: Millikan on concepts and teleosemantics; Cummins, McGinn, Isaac, Craik, Ramsey, Swoyer, Churchland, and Waskan on isomorphism and psychosemantics; Dretske, Jacob, Neander, and Shea on information (vs. isomorphism); Dretske and Prinz on incipient causes and Sainsbury and Tye on originalism; Fodor on triangulation; Fodor on triangulation, Rupert on best test, and Eliasmith & Usher on statistical dependence; Papineau on success semantics; Dennett on the intentional stance; psychological theories of concepts and concept acquisition; direct reference; embodied cognition; mental files; and psychological essentialism.

Chapter 12: An empirical reconception of the concept of mind

Our concept of intentionality has its home within our biological species, but the kind that the concept picks out likely extends beyond this. There are structures homologous to human cortex throughout mammals, reptiles, birds, fish, and even arthropods (the mushroom bodies), and they may operate on similar design principles. Further, these principles may be a natural, stable endpoint of many possible selectional processes, in which case our kind of intentionality may not be merely terrestrial. On the other hand, empirical inquiry may reveal that our cortical design principles occupy a tiny region in a vast space of continuous variation, like visible light in the spectrum of electromagnetism.

The underlying nature of our intentionality may be discovered only by answering empirical questions. Hopefully the way forward is now clearer.

(This chapter also contains a retrospective of the book.)

References:

Becker, S., & Hinton, G. E. (1992). Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, *355*, 161-163.

Dretske, F. (1988). *Explaining Behavior*. Cambridge, Mass.: MIT Press.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A theory of causal learning in children: causal maps and Bayes nets. *Psychological review*, *111*(1), 3.

Grossberg, S. (2013). Adaptive Resonance Theory: How a brain learns to consciously attend, learn, and recognize a changing world. *Neural Networks*, *37*, 1-47.

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences*, *27*, 377-442.

Phillips, W. A., & Singer, W. (1997). In search of common foundations for cortical computation. *Behavioral and Brain Sciences*, *20*, 657-722.

Werning, M. (2003). Synchrony and composition: Toward a cognitive architecture between classicism and connectionism. In *Foundations of the Formal Sciences II* (pp. 261-278). Springer.

Werning, M., & Maye, A. (2007). The cortical implementation of complex attribute and substance concepts: Synchrony, frames, and hierarchical binding. *Chaos and complexity letters*, *2*(2).