David Danks – Unifying the Mind: Cognitive Representations as Graphical Models
MIT Press, 2014, ISBN: 978-0-262-02799-1

In *Unifying the Mind*, David Danks aims to equip the advocate of mental representations with a further weapon by developing a unifying framework based on the existence of a common store of representations, structured as graphical models, which underlie a number of cognitive processes. Danks offers two initial motivations for the existence of a unified account, (1) the need to explain the relatively seamless manner in which we seem to shift between distinct cognitive operations, and (2) our ability to attend to goal-relevant information [p.4].

Though well established in machine learning, the notion of a graphical model may be new to those in other areas of the cognitive sciences. Graphical models are probabilistic representations of relevance relations, which have a qualitative component (the graph), as well as a quantitative component (the formalism). The graph consists of a number of nodes and connecting edges, both of which have different meanings depending on the role they are posited to fulfil, as well as the type of graphical model. This variety, Danks argues, is what makes graphical models flexible enough to deal with a wide range of cognitive processes, and allows him to show how this novel cognitive architecture can unify previously isolated areas of cognition.

Unification is a theoretical virtue highlighted by several other notable frameworks that seem to be growing in popularity (one of which is discussed in section 8.2.2 when Danks discusses the Bayesian approach to modelling cognition). However, Danks' framework differs in a number of ways from these alternatives, making it a worthwhile read even for those who support another framework.

Firstly, Danks is careful to acknowledge the limited scope for the framework, when he states early on that it is not the case that the graphical models framework can trivially accommodate or represent all theories in cognitive science. This is made clear in section 4.4, when Danks provides an account of causal perception that he argues is unlikely to be accommodated within the graphical models framework. Rather than being seen as a limitation of the entire framework, this may be comforting to those familiar with charges of unfalsifiability levied against other unifying frameworks (cf. Bowers & Davis, 2012).

Secondly, in addition to being immensely clear both in style and structure, Danks manages to present a rigorous conceptual analysis of the framework without the requirement to engage with the mathematical details. Though the formal details are presented (Chapter 3), sections marked with an asterisk highlight the mathematical treatments, and can be skipped by those who do not wish to engage with the formalism. This was a useful decision, and those wishing to introduce these concepts in a graduate level course will benefit from the consideration that has gone into both the structure and presentation of this book.

Danks begins by situating the framework in context, and highlights the importance of the notion of relevance that is explored more fully in later chapters. A key insight of the framework is that the *graph* aspect of a graphical model encodes relevance relations in a compact manner that makes such representations easy to store and employ, and may help to provide explanations for certain aspects of learning and decision-making (e.g. our ability to attend to goal-relevant information).

This is succeeded by an important philosophical treatment of the framework. To assist with this discussion, in section 2.1 Danks provides a useful exposition of the differences between a *model*, *architecture* and *framework* (in terms of their relative abstractness):

- "[A] cognitive *model* offers a computationally well-specified account of a specific aspect of cognition." [p.14]
- "[A] cognitive *architecture* specifies the basic building blocks for cognitive models, as well as the ways in which those components can be assembled." [p.14]
- "Most abstractly, a cognitive *framework* is a general approach for understanding the nature of the mind." [p.15]

Danks claims that his cognitive architecture is based on the existence of shared representations structured as graphical models, and is therefore committed to specific computational models. He also comments on the initial absence of a cognitive *theory*, stating that the term is employed in a number of ways to refer to any of the aforementioned notions - its usage is reserved for when distinctions between levels is not important. The three levels outlined are employed carefully throughout, and are useful distinctions, especially in later chapters (7-9) when Danks turns to considering the architectural commitments of the framework in relation to alternative architectures.

For those from a philosophy of science background, much of this material will be familiar, but for those unfamiliar with the discussion, this is an important discussion that sets the stage for the following chapters. Most notably, it provides an important foundation for the architecture's metaphysical commitments, including why it is committed to *representational realism*, but not *process realism.* Although chapters 4-6 present empirical evidence to support the graphical models framework, as Danks acknowledges, the fact that a common formalism can model multiple cognitive phenomena does not entail a realist commitment to a particular cognitive representation. Instead, Danks wishes to make the more substantive claim that the cognitive architecture he posits is in fact the *best explanation* for a range of empirical data, including data from cross-cognition transfer studies.

This is achieved most clearly at a later stage in the work (Chapter 7), when Danks explores a "possibility space" of cognitive architectures, which vary in the amount of sharing of *representations* between what Danks calls "silos" (not to be confused with modules). The idea of a silo is taken to refer to a cognitive *process* (e.g. causal learning or feature inference), whereby a particular silo may access a single store or multiple stores of representations for any given task, and individual representations may be shared across multiple silos. How coarsely grained we choose to individuate these silos is understandably important, and is discussed in this chapter along with additional nuances. Experiments that test whether any substantive cross-cognition transfer occurs, should be able to discriminate between multiple-store and single-store accounts. Danks discusses some extant experiments, as well as putting forward some novel experiments that could be performed.

Danks claims that within his framework "large swaths of human cognitive activity can be understood as different operations on a shared representational store" [p.151]. In terms of the *possibility space* of different architectures, the graphical models architecture, therefore, posits a *single* representational store, and *multiple* cognitive processes or silos. As such, Danks argues that his framework is committed to *representational realism*, without making any commitments to *process realism*.

In addition to the empirical data explored in the discussion of cross-cognition transfer, the middle sections of the book discusses three specific cognitive processes in greater detail. Causal cognition is explored in chapter 4; familiar notions of concepts and categories in cognitive psychology are discussed in chapter 5, and finally, chapter 6 explores how causal knowledge can be applied to decision-making. In these chapters Danks covers a lot of ground, and some of the treatments may be too quick for those less familiar with some of the extant literature. Arguably, the extensive bibliography serves the function of pointing the interested reader in the right direction, and as this

book is intended for an interdisciplinary audience, perhaps this decision was wise in order to avoid bloat. However, I found myself giving Danks the benefit of the doubt on a number of occasions that the graphical models did in fact adequately capture the empirical data.

One challenge that is raised in these middle chapters is worth mentioning in further detail. In sections 4.3 and 5.3, Danks focuses on two noteworthy empirical studies (Rehder & Burnett, 2005; Rottman & Hastie, 2014). These studies focus on a subjects' behaviour as a violation of the Markov assumption (a fundamental part of the graphical models framework), which states that a non-adjacent node should be informationally independent (irrelevant) to the node of interest when conditioning on its adjacent nodes. Such a criticism is obviously a salient objection for Danks' framework. However, Danks provides two responses to these studies. The first focuses on implicit causal relevance relations that are not part of the initial models, but nevertheless provide "natural explanations" of the phenomena [p.111]. The second focuses on a subjects' inherent uncertainty, or possibility of error, in causal inference or feature identification, which Danks claims can lead to an increase in the amount of relevant information that may be gleaned from seemingly independent properties or events [pp.112-3]. These responses help to further highlight the importance of the exposition outlined in chapter 2. The "natural explanations" Danks appeals to demonstrate how the *models* posited are constrained by other domains of knowledge such as folk psychology and folk biology, and so rather than merely fitting the data, these responses aim to go further and consider a wider explanatory scope.

Danks concludes by turning to some broader implications (rethinking the notion of modularity), and remaining challenges for the framework (capturing multiple relevance relations). He also notes the absence of neuroscientific data, stating that current neuroimaging data is simply too coarse grained to be directly relevant to the structural details raised by his cognitive architecture. Though Danks may be right at present, there is emerging research that challenges the very computational perspective that Danks defends (Anderson, 2014). This alternative picture explores an embodied approach to cognition, which sees the brain as a deeply interactive control mechanism for a situated agent, and is supported by large scale meta-analyses of neuroimaging data as well as wide-reaching areas of the cognitive sciences. Such an action-oriented picture of the mind conflicts with Danks' "unabashedly computational" perspective (see chapters 5&6 of Ibid.), but as of yet it does not have as well-specified an architecture as the one presented by Danks. Although chapter 6 explores how to understand the process of decision-making within the graphical models framework, one area where it could benefit enormously from is in providing a more detailed account of an agents' interactions with its environment, and how this is supported by the representations of the graphical models architecture. This is something Danks acknowledges when he addresses the absence of the study of language and social cognition in chapter 10. Such a development will undoubtedly strengthen the framework, and widen its explanatory and unificatory scope.

There is no doubt that Danks is knowledgeable of a wide range of interdisciplinary work in the cognitive sciences, and his ability to cogently present the material is commendable. This is an instructive and thought-provoking book, but more than that, it is an excellent example of how philosophical work in the cognitive sciences should be done.

References

Anderson, M. L. (2014). *After phrenology: Neural reuse and the interactive brain*. MIT Press.

Bowers, J. S., & Davis, C. J. (2012). Bayesian just-so stories in psychology and neuroscience. *Psychological bulletin*, *138*(3), 389.

Rehder, B., & Burnett, R. C. (2005). Feature inference and the causal structure of categories. *Cognitive Psychology*, *50*(3), 264-314.

Rottman, B. M., & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological bulletin*, *140*(1), 109.