A Computational Foundation for the Study of Cognition*

David J. Chalmers

Australian National University
New York University
chalmers@anu.edu.au

Computation is central to the foundations of modern cognitive science, but its role is controversial. Questions about computation abound: What is it for a physical system to implement a computation? Is computation sufficient for thought? What is the role of computation in a theory of cognition? What is the relation between different sorts of computational theory, such as connectionism and symbolic computation? In this paper I develop a systematic framework that addresses all of these questions.

Justifying the role of computation requires analysis of *implementation*, the nexus between abstract computations and concrete physical systems. I give such an analysis, based on the idea that a system implements a computation if the causal structure of the system mirrors the formal structure of the computation. This account can be used to justify the central commitments of artificial intelligence and computational cognitive science: the thesis of computational sufficiency, which holds that the right kind of computational structure suffices for the possession of a mind, and the thesis of computational explanation, which holds that computation provides a general framework for the explanation of cognitive processes. The theses are consequences of the facts that (a) computation can specify general patterns of causal organization, and (b) mentality is an *organizational invariant*, rooted in such patterns. Along the way I answer various challenges to the computationalist position, such as those put

* This paper was written in 1993 but never published (although section 2 was included in "On Implementing a Computation", published in *Minds and Machines* in 1994). Because the paper has been widely cited over the years, I have not made any changes to it apart from adding one footnote, instead saving any further thoughts for my reply to commentators. In any case I am still largely sympathetic with the views expressed here, in broad outline if not in every detail.

Journal of Cognitive Science 12: 323-357, 2011

forward by Searle. I close by advocating a kind of minimal computationalism, compatible with a very wide variety of empirical approaches to the mind. This allows computation to serve as a true foundation for cognitive science.

Key words: *computation; cognition; implementation; explanation; connectionism; computationalism; representation; artificial intelligence*

1. Introduction

Perhaps no concept is more central to the foundations of modern cognitive science than that of computation. The ambitions of artificial intelligence rest on a computational framework, and in other areas of cognitive science, models of cognitive processes are most frequently cast in computational terms. The foundational role of computation can be expressed in two basic theses. First, underlying the belief in the possibility of artificial intelligence there is a thesis of *computational sufficiency*, stating that the right kind of computational structure suffices for the possession of a mind, and for the possession of a wide variety of mental properties. Second, facilitating the progress of cognitive science more generally there is a thesis of *computational explanation*, stating that computation provides a general framework for the explanation of cognitive processes and of behavior.

These theses are widely held within cognitive science, but they are quite controversial. Some have questioned the thesis of computational sufficiency, arguing that certain human abilities could never be duplicated computationally (Dreyfus 1974; Penrose 1989), or that even if a computation could duplicate human abilities, instantiating the relevant computation would not suffice for the possession of a mind (Searle 1980). Others have questioned the thesis of computational explanation, arguing that computation provides an inappropriate framework for the explanation of cognitive processes (Edelman 1989; Gibson 1979), or even that computational descriptions of a system are vacuous (Searle 1990, 1991).

Advocates of computational cognitive science have done their best to repel these negative critiques, but the positive justification for the foundational theses remains murky at best. Why should *computation*, rather than some other technical notion, play this foundational role? And why should