Foreword

Frank E. Ritter, Dec 30, 2015

This book describes the background and results from Jun Tani's work of over a decade of building robots that think and learn through interaction with the world. It has numerous useful and deep lessons for modelers developing and using symbolic, subsymbolic, and hybrid architectures, so I am pleased to see it in the *Oxford Series on Cognitive Models and Architectures*. It is work that is in the spirit of Newell and Simon's (1975) theory of empirical exploration of computer science topics and their work on generation of behavior, and also takes Newell and Simon's and Feynman's motto of understanding through generation of behavior seriously. At the same time, this work extends the physical symbol hypothesis in a very useful way by suggesting by example that the symbols of human cognition need not be discrete symbols manually fed into computers (which we have often done in symbolic cognitive architectures), but can instead be composable neuro-dynamic structures arising through iterative learning of perceptual experience with the physical world.

Tani's work has explored some of the deep issues in embodied cognition, about how interaction with the environment happens, what this means for representation and learning, and how more complex behavior can be created or how it arises through more simple aspects. These lessons include insights about the role of interaction with the environment, consciousness and free will, and lessons about how to build neural net architectures to drive behavior in robots.

The book starts with a review of the foundations of this work, including some of the philosophical foundations in this area (including the symbol grounding problem, phenomenology, and the role of time in thinking). It argues for a role of hierarchy in modeling cognition, and for modeling and understanding interaction with an external world. The book also notes that state space attractors can be a useful concept in understanding cognition, and, I would add, this could be a useful additional way to measure fit of a model to behavior. This review also reminds us of areas that current symbolic models have been uninformed by—I don't think that these topics have been so much ignored as much as put on a list for later work. These aspects are becoming more timely, as Tani's work shows they can be. The review chapters make this book particularly useful as an advanced textbook, which Tani already uses it for.

Perhaps more importantly, in the second half of the book (Chapters 6 to 11) Tani describes lessons from his own work. This work argues that behavior is not always programmed or extant in a system, but that it can or often should arise in systems attempting to achieve homeostasis—that there are positions of stability in a mental representation (including modeling others, imitation), and that differences in knowledge between the levels can give rise to effects that might be seen to be a type of consciousness, a mental trace of what lower levels should do or are doing, or explanations of what they have done based on predictions of the agent's own behavior, a type of self-reflexive mental model. These results suggest that more models should model homeostasis and include more goals and knowledge about how to achieve it.

His work provides another way of representing and generating behavior. This way emphasizes the dynamic behavior of systems rather than the data structures used in more traditional approaches. The simple ideas of evolution of knowledge, feedback, attractors, and further concepts provide food for thought for all systems that generate behavior. These components are reviewed in the first part of the book. The second part of the book also presents several systems used to explore these ideas.

Lessons from this book could and should change how we see all kinds of cognitive architectures. Many of these concepts have not yet been noticed in symbolic architectures, but they probably exist in them This new way to examine behavior in architectures has provided insights already about learning and interaction and consciousness. Using these concepts in existing architectures and models will provide new insights

into how compositional thoughts and actions can be generated without facing the notorious problems of the symbol grounding problem or, ultimately, the mind–body problem.

In his work about layers of representation, he has seen that higher levels might not just lead the lower levels, but also follow them, adjusting their own settings based on the lower levels' behavior. An interpretation of the higher levels trying to follow or predict the lower levels provides a potential computational description and explanation of some forms of consciousness and free will. I found these concepts particularly intriguing. Not only that higher levels could follow and not lead lower levels, but that the mismatch could lead to a kind of postdiction in which intention becomes consciously aware after action. We might see this elsewhere as other architectures, their environments, and their interaction with the environment become more complex, and indeed should look for it.

I hope you find the book as useful and suggestive of new areas of work and new aspects of behavior to consider for including in architectures as I have.

Contents