Can we model the mind with its own products?

Marcin Chady
The University of Birmingham, UK
M.Chady@cs.bham.ac.uk

Abstract

This poster is an invitation to a free discussion on the limitations of the mind in its capability to understand itself and other minds. It poses some questions, presents examples and makes suggestions, in the hope to attract some comments and feedback from the readers.

The human brain has evolved to model the world, so that we can avoid dangers and achieve our goals. It is not possible to simulate the universe exactly, so some (arguably significant) degree of approximation is required. In order to draw conclusions and predict future events, one must generalise, so any brain-endowed organism is a natural classifier, and we humans are a prime example of this: we categorise, generalise and organise things into hierarchies, so that we can make better sense of the world, and turn its rules into our advantage. We have developed languages, algebras and calculi. Our sophisticated system of symbols can describe abstract as well as physical things. However, as we face the enigma of our own minds, we must ask ourselves the question: Can we model the mind with its own products?

When we talk about thinking, we tend to use inherently ambiguous terms: emotions, intentions, exploration, control, or even memory. It isn't until we actually try to realise them in a physical system, that it transpires how vague they are. Many of them have been discredited from scientific discourse, and confined to the realms of folk psychology, but much of the confusion remains (Smolensky (1988)). People like to view the mind as a number of parallel processes: transforming information, exchanging signals, making decisions, each specialised in a different kind of operations. It is easy to conceptualise such systems using the "divide and conquer" strategy, and, perhaps most importantly, it is easy to *depict* them using boxes and arrows.

Although symbolic systems allow a certain amount of flexibility and/or indeterminism in their architecture (cf. the ACT-R system or Sloman's models of emotional agents, see e.g. Sloman (1999)), the component-based perspective still remains. It is easy to forget that the functionality and the mechanism are not necessarily in a one to one relationship. Consequently, we tend to make statements like "if there are two motive generators competing, then there must be a motive comparator". Such reasoning assumes that each function identified in the designer's mind has an architectural embodiment. Yet, it is easy to imagine competition without an arbiter, in which case the "comparator" has no counterpart. And what about all those possibilities which are not so easy, or even impossible to imagine?

Much of the non-symbolic AI, such as neural networks, inadvertently follows the same path by giving parts of their systems explicitly defined functionality, and combining them into interacting ensembles of modules. A prime example of this are implementationalist systems, where the logic is hand-wired into the connections between units (e.g. Shastri and Ajjanagadde (1993), Barnden (1991)). Not only do they risk missing out a significant part of the solution domain, but parts of the problem domain too: they usually end up with a limited capability for learning.

It is hard not to structuralise, given that this is what our brains have been designed to do through millennia of evolution. However, one must not forget that everything our cognitive processes come up with is an artefact. It's a product of a process whose purpose is by no means general. On the contrary, its purpose is to filter out details which are irrelevant to survival, and produce a *convenient* model of the environment. There is no reason why our cognitive apparatus should be able to cope with its own workings, just like no one is predisposed to imagine 7-dimensional objects, which some physicists suggest our universe really consists of.

Can we escape the kaleidoscope of our cognition? Most probably not. Even the machines we create will be biased. However, we can reduce this bias by limiting our involvement in their design, putting emphasis on self-organisation rather than designing complete architectural solutions. Evolutionary strategies spring to mind. But we can also study complex unconventional systems and look for new insights there. Of course, every insight will be subjective too, but there is nothing we can do about it. At least we may discover a new way of thinking which will take us a further step back from the narrow vision of a "survival machine".

An example of such approach can be found in the

work of Hanson and Crutchfield (1992). There, a complex behaviour of CA is analysed using classic paradigms from computational theory, such as Finite State Automata. The work demonstrates how complicated operation can result from simple interactions between simple units. Conversely, the use of classical computational models to describe the global behaviour shows that no explicit FSA machinery is necessary to produce FSA behaviour.

Perhaps then, there are no parts responsible for motivation, emotions, control, etc. They could be just facets of the same process, presenting themselves differently from different angles. Much like the facets of a hypercube intersected with a 3-dimensional hyperplane, or like ephemeral vortices of turbulent flow. And, last but not least, let us not forget that the mind itself is an artefact: an attractive and convenient concept for picturing our thoughts perhaps, but not necessarily the most useful one in tackling the "hard problem".

References

John A. Barnden. Encoding complex symbolic data structures with some unusual connectionist techniques. In John A. Barnden and J. B. Pollack, editors, *High Level Connectionist Models*, volume 1 of *Advances in Connectionist and Neural Computation Theory*. Ablex, Norwood, N.J., 1991.

Jim Hanson and Jim Crutchfield. The attractor-basin portrait of a cellular automaton. *Journal of Statistical Physics*, 66:1415–1462, 1992.

Lokendra Shastri and Venkat Ajjanagadde. From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences*, 16:417–494, 1993.

Aaron Sloman. What sort of architecture is required for a human-like agent? In Michael Wooldridge and Anand Rao, editors, *Foundations of Rational Agency*. Kluver Academic, Dordrecht, 1999.

Paul Smolensky. On the proper treatment of connectionism. *The Behavioral and Brain Sciences*, 11, 1988.