Design and implementation of a biologically inspired agent architecture combining emotions and reasoning

Janos Tapolczai
Thesis supervisor: A.o. Univ.-Prof. Dr. Hans Tompits

March 27, 2015

1 Problem statement

To date, emotions in AI have very much occupied a niche; instead, problems like search, reasoning, pattern recognition and statistics have dominated the field. People like Marvin Minsky and Aaron Sloman, however, have argued that emotions play key roles in cognition — to wit, Minsky's famous saying [12, p. 163]:

> The question is not whether intelligent machines can have any emotions, but whether machines can be intelligent without any emotions.

One of the major unsolved problems — in biology as well as AI — is the interaction between emotions and reasoning: how do they help and how do they hinder the reasoning process? Could reasoning function without emotions? Are they just an external influence, or an integral part of the reasoning process?

This thesis would build primarily on Minsky's [12, 13] and Sloman's work [18], who seek to answer these questions with the help of evolutionary neuroscience: they only consider architectures that could have evolved gradually. By thus constraining their models through the criterion of evolvability, they sketch plausible histories through the which the human brains might have gone, accruing one subsystem at a time. The goal of this work would be the application of their method to affect and reasoning specifically, and to the development of a cognitive architecture that incorporates both.

2 Expected results

The primary aim is the creation of an affective toy-AI that has to navigate a simple, competitive world and which does not perform reasoning as such, but uses emotions and rudimentary prediction of the future to choose its actions. Its architecture would be hard-wired, but it would be parametrisable via its emotional reactions to stimuli: whereas one agent might react to the sight of an another agent with anger, a differently parametrised agent might react with fear.

The expectation is that these agents will be able to effectively navigate said toy world which, to a rough approximation, resembles the typical environment of animals: competition for food, cooperation and conflict with other, similar agents, and reproduction. If the toy AI and its underlying architecture work as intended, we should see patterns of behaviour similar to those we find in real organisms: some would be overly frightful, fleeing from every threat; some would react aggressively to almost anything; yet others would act out an iterated prisoner's dilemma in which they would slowly become more cooperative with agents with whom they had positive interactions in the past, yet would quickly develop a lasting aversion if a negative interaction occurred.

This test scenario would (a) determine how successful the overall model is and (b) how successful different parameter sets w.r.t. emotional reactions within the confines of the model are. If at least some subset of agents can keep itself from extinction, we would have an indication that their behaviour at least matches that of some lower animal in terms of effectiveness.

In case of a success, we would have a proof-of-concept that the evolutionary approach is workable and useful in the design of AIs, and that this sort of "neurological archaeology" or constructing cognitive architectures by analysing evolutionary steps of real brains has its place besides traditional approaches like neural networks and symbolic computation.

3 Methodology

The thesis would be split into two parts: a theoretical and a practical one. The first, theoretical part would analyse nervous systems and brain structures; the second, practical part propose a cognitive architecture based on the first and on the work of Aaron Sloman. The basis of the entire work would be an evolutionary model: with the help of biology, we would look at each subsystem in the brain — how it developed, on what other subsystem it relied, what analogues exist in non-human species, how it interacts with other subsystems — and consequently develop a cognitive architecture.

The underlying hypothesis is that, since brains are evolved organs, their functioning can best be understood through re-tracing their development through various stages. By requiring each development in brain function to be relatively simple and useful by itself, we can constrain the space of possible designs. This approach of design space and (evolutionary) niche space was, as said above, proposed by Aaron Sloman [18] and is a running theme in many of his works [20, 17, 19].

The interaction between affect and reasoning being the focus of this work, re-tracing brain development would involve giving an account of the following:

1. the structure and purpose of affect;
2. the purpose and mechanism of planning; and
3. the structure of reasoning, as it relates to the above two, and as humans actually perform it.

The second, practical part, would serve as a test for the developed model. Here, we use simulation as a way of evaluation. The aim is to develop an AI that, in simple model worlds, can perform with comparable effectiveness as animals do in the real one. Hence, the agents would be tested in a toy world that poses the typical challenges that animals face (in an extremely simplified form, of course):

1. hunger, which necessitates the gathering of resources,
2. predators (agents with a simple "flight-or-fight" AI),
3. plants that can be eaten, and
4. other, similar agents, with which one can interact in hostile or cooperative ways.

Successful agents would reproduce and their offspring would inherit their emotional profiles. To allow agents with different personalities to better differentiate themselves from others, the protocol of interaction would be minimalistic: an agent can outright attack another one, give it some food item, or send a gesture (a string) of its choosing. It would be incumbent upon the recipient of such gestures to interpret them in accordance with its emotional profile — there would be no a priori knowledge about which gestures meant what; two different agents could interpret the same signal in completely different ways.

4 State of the art

Research into human-level AI is proceeding slowly. Though AIs can outperform humans at a variety of tasks (notably games like chess and bridge [9] that are amenable to standard search techniques like A* and IDDFS, but also complex, real-world problems like driving cars [7, 16] or reasoning [10, 6]), an artificial intelligence that can compete with a human at many different tasks has yet to be created. While most effort today goes into the creation of such weak AIs, there are a few influential people — Marvin Minsky, John McCarthy, Nils Nilsson, and Patrick Winston, among others [15, p. 27] — who wish to create programs that act in a believably human-like way.

On the other hand, much research has been done in the field of affective neuroscience and the basis of emotions (for an overview, see *The Cambridge Handbook of Human Affective Neuroscience* [1]). These findings have been applied to artificial intelligence to some degree — examples being Cynthia Breazeal's work at MIT [3, 2], and the emotion-driven robots of Sandra Gadanho and John Hallam [8].

A number of cognitive architectures do exist, from the purely theoretical [13] to those actually implemented, such as CMU's 4CAPS [21] and those based on R. A. Brooks' subsumption architecture [4]. In *Intelligence without Reason* [5], Brooks does, in fact, come quite close to our goal: a biologically inspired architecture that does not have a priori knowledge about reasoning, but merely synthesizes it from the interplay of simpler components.

5 Context

This thesis deals with affective AI; as such, it relates very closely to the curriculum of Computational Intelligence.

References

[1] Jorge Amory and Patrik Vuilleumier. *The Cambridge Handbook of Human Affective Neuroscience*. Cambridge University Press, 2013.

[2] Cynthia Breazeal. Emotion and sociable humanoid robots. *Internation Journal of Human-Computer Studies*, 59:119–155, 2003.

[3] Cynthia Breazeal. Kismet. `http://www.ai.mit.edu/projects/sociable/overview.html`, 11 2014.

[4] Rodney A. Brooks. A robust layered control system for a mobile robot. *Robotics and Automation, IEEE Journal of*, 2(1):14–23, Mar 1986.

[5] Rodney A. Brooks. Intelligence without reason. In *Computers and Thought, IJCAI-91*, pages 569–595. Morgan Kaufmann, 1991.

[6] Michael Fink, Stefano Germano, Giovambattista Ianni, Christoph Redl, and Peter Schüller. ActHEX: Implementing hex programs with action atoms. In Pedro Cabalar and TranCao Son, editors, *Logic Programming and Nonmonotonic Reasoning*, volume 8148 of *Lecture Notes in Computer Science*, pages 317–322. Springer Berlin Heidelberg, 2013.

[7] Adam Fisher. Inside Google's Quest To Popularize Self-Driving Cars. `http://www.popsci.com/cars/article/2013-09/google-self-driving-car`, 9 2013.

[8] Sandra Clara Gadanho and John Hallam. Robot Learning Driven by Emotions. *Adaptive Behaviour*, 9(1):42–64, 2001.

[9] Jonathan Jonathan Schaeffer. The Games Computers (and People) Play. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on on Innovative Applications of Artificial Intelligence, July 30 - August 3, 2000, Austin, Texas, USA.*, page 1179. AAAI Press / The MIT Press, 2000.

[10] Vladimir Lifschitz. Action Languages, Answer Sets and Planning. In *In The Logic Programming Paradigm: a 25-Year Perspective*, pages 357–373. Springer Verlag, 1999.

[11] John McCarthy. From Here to Human-level AI. *Artif. Intell.*, 171(18):1174–1182, December 2007.

[12] Marvin Minsky. *The Society of Mind*. Simon & Schuster, New York, 1988.

[13] Marvin Minsky. *The Emotion Machine*. Simon & Schuster, New York, 2006.

[14] Marvin Minsky, Push Singh, and Aaron Sloman. The st. thomas common sense symposium: Designing architectures for human-level intelligence. *AI Mag.*, 25(2):113–124, June 2004.

[15] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, New Jersey, 2010.

[16] Tom Simonite. Data Shows Google's Robot Care Are Smoother, Safer Drivers Than You or I. http://www.technologyreview.com/news/520746/data-shows-googles-robot-cars-are-smoother-safer-drivers-than-you-or-i/, 10 2013.

[17] Aaron Sloman. The Mind as a Control System. *Royal Institute of Philosophy Supplement*, 34:69–110, 3 1993.

[18] Aaron Sloman. Exploring Design Space and Niche Space. In *In Proceedings 5th Scandinavian Conference on AI*. IOS Press, 1995.

[19] Aaron Sloman. What Sort of Control System Is Able to Have a Personality? In *Creating Personalities for Synthetic Actors, Towards Autonomous Personality Agents*, pages 166–208, London, UK, UK, 1997. Springer-Verlag.

[20] Aaron Sloman. Beyond shallow models of emotion. In *Cognitive Processing: International Quarterly of Cognitive Science*, pages 177–198, 2001.

[21] Carnegie-Mellon University. 4CAPS Cognitive Neuroarchitecture. http://www.ccbi.cmu.edu/4CAPS/index.html, 11 2014.