Procedia Computer Science
Volume 88, 2016, Pages 155–162
7th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2016

Evaluation of Cognitive Architectures Inspired by Cognitive Biases

Christoph Doell[1] and Sophie Siebert[1]

Otto von Guericke University of Magdeburg, Germany
doell@ovgu.de sophie.siebert@st.ovgu.de

Abstract
Cognitive architectures are frequently built to model naturally intelligent behavior. This aims on two primary goals: On one hand these architectures model human behavior in order to give a better understanding of the human thought process. On the other hand cognitive architectures are an approach of modeling artificial intelligence. Those two goals might be conflicting, as humans sometimes act irrationally e.g. because they were cognitively biased. In this work, we analyze on a theoretical level whether cognitive architectures are also biased. Therefore we first abstract more general behavior from cognitive fallacies. Then we evaluate for the architectures Clarion, Leabra and Lida to what extent they can be biased.

Keywords: Cognitive Fallacy, Evaluation, Cognitive Architecture

1 Introduction

Most cognitive architectures are built to model specific aspects of human cognition and therefore they are evaluated in those tasks they were created for: The Chrest architecture[1] aims on modeling human perception. Therefore it successfully simulates humans eye movements, when playing chess. In contrast, the cognitive architecture 'Adaptive Control of Thought Rational' (ACT-R) [2] models the human thought process. It can for example simulate the processing of cognitive arithmetic [3]. The BICA Society gives an overview showing 26 cognitive architectures [4] with their different goals and approaches. Comparing these architectures is difficult, as they process different kinds of data in different ways. Therefore questions like 'Which of them model human cognition the best?' cannot be clearly answered. In this work, we propose to compare cognitive architectures, by a general measure, that can be applied even if architectures are built to process different kinds of inputs. The key to this measure is cognitive biases.

When solving tasks, humans do not always act rationally. They tend to rely on learned heuristics and make systematic errors – known as cognitive biases, e.g.:

- When humans were asked to estimate their own skill, Kruger and Dunning [5] showed that experts tend to underestimating, while unskilled individuals tend to overestimating.

Selection and peer-review under responsibility of the Scientific Programme Committee of BICA 2016

doi:10.1016/j.procs.2016.07.419

- When asked about the valuation on furniture, self-made products achieve higher values, even if all the parts were bought [6]. This is known as *IKEA effect*.
- People were asked two questions in a row, both on estimating a number. First, they had to decide whether or not the estimated number was bigger than a given value. Then they were asked for the exact value. The results showed, that the given value highly influenced the exact estimation; the answered value was close to the anchoring value[7].

Many of these effects are known, but they mostly apply only to a specific given task. Thus, they can hardly be applied directly to evaluate cognitive architectures. This is why we generalize them in order to make them applicable for a measurement of cognitive architectures.

Cognitive fallacies are often judged as weakness in humans behavior, as the actions are considered irrational. In contrast to this opinion Gigerenzer [8] showed evidence that in real life situations, with limited knowledge, human behavior often leads to better, more robust results, than rationally acting algorithms. Following this argumentation, it is possible, that those very processes which cause cognitive fallacies or biases also result in general intelligent behavior.

Given a natural cognitive system as inspiration for the construction of an artificial cognitive architecture. When modeling, the focus can lie on the functionality or the structure of the original. Functionality is important for the performance of the system, while the structure might give hints on the internal processes, and possibly lead to better explanations[9].
Thus, this work investigates how structural differences can lead to functional differences. Further it builds a theoretical basis for later empirical studies to compare cognitive architectures. We analyze the internal processes of three exemplary cognitive architectures to predict whether they can be biased. In the following section, we give more details on cognitive biases and generalize the behavior in order to be able to measure the effect in a cognitive architecture. Therefore we summarize the effects of cognitive biases abstractly. In Section 3 we describe three exemplary cognitive architectures: Clarion, Lida and Leabra. Their behavior with respect of the abstracted features are explained in Section 4. Finally, in Section 5, the results are summarized and an outlook on future empirical evaluations is given.

2 Cognitive Biases

Cognitive biases are effects indicating a systematic deviation from a rational judgment. In an early work, Kahneman and Tversky [7] presented 13 different cognitive biases. For example they provided evidence that human's natural way of decision making ignores prior probabilities. The *availability heuristic* shows up when estimating probabilities of occurrences of events. Humans tend to estimate the ease of retrievability, instead.

Participants were asked to decide for one of two mathematically identical alternatives – one was described with mostly positive words, the other one used negative words[10]. People significantly preferred the positive formulation – the corresponding effect is called *framing effect*.

Another long known psychological effect is the *halo effect*. Few attributes of a person were read to participants of a study. When the participants were later asked to characterize the person, the result was highly influenced by the *order* of the given attributes [11].

Another study described the *priming effect*: people ware asked to complete the letters *'so_p'* such that an existing word is created. The result varied: When people had the context of cleaning or washing in mind, the word they came up with was *soap*. When the context was set to food, people answered *soup*[12, p. 72].

Abstracting from these examples, for a cognitive bias to be visible, we need a system with three properties: