How suited are cognitive architectures for implementing moral reasoning? – a Systematic Literature Review

TU Delft
by Wojciech Hajdas

Background

Problem:

Artificial Agents are used in an increasing variety of domains with moral and ethical consequence including:

- Healthcare support systems
- Automation of creation of legal documents
- Educational technologies

Thus, researchers look for effective ways of implementing understandable ethical consideration into software.

Analyzed Solution:

Some researchers use cognitive architectures – digital systems that aim to model human cognition and behavior, to implement moral reasoning – evaluating and justifying of actions in terms of right/wrong [1].

Why cognitive architectures?

In theory, due to their human-likeness, cognitive architectures (Cas) offer better psychological plausibility and explainability, as well as offer support for other useful, cognitive functions such as norm acquisition and adaptation over time[2], in contrast to their non-biologically inspired counterparts. These factors potentially play a role in the overall success of the implemented systems.

Reasons for my research?

There exist no comprehensive survey of current practical implementations of moral reasoning using cognitive architectures, which could be helpful for future research .

Research question

How suited are cognitive architectures for implementing of the moral reasoning, based on current implementations in the literature?

Suitability - subjective, composite measure composed of:
- scale and capabilities
- challenges during development
- overall researchers' attitudes

Also meant to provide overview of the field.

Search Strategy Methods

All database calls are (approx.):
(“cognitive architectur*” OR “ACT-R” OR “SOAR” OR “LIDA) AND (“ethic*” OR “moral*”)

9 studies included in the data extraction process

Methodology

Quotes extracted into data extraction table once manually and once with the use of Google LM.

Selected quotes then turned into answers of specific “extraction” questions

Results

2 research paradigms

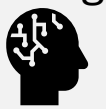
Human modeling
Artificial Cognitive Agents

Fragmentation of the field

Small Scope and Capabilities

Problems with validation

Correct reasoning and inner workings



Conclusion

Cognitive architectures are conditionally suitable for implementing of moral reasoning. They do well in tasks that require modelling of human reasoning, and they exhibit their desired behaviors on small scale, unrealistic simulations.

They are a complex tool, used for implementing of moral reasoning – also a complex task. This combination might be the main cause of the rather unimpressive results, even despite the possible validity of the approach.

Future Work

Systematic Literature Review analyzing proposals for the ethical guidelines of autonomous systems.

↓

Development of benchmarks and standards

↓

Comparative studies of CAs

References

[1] Reynolds, K. (2023). Moral reasoning. In *Springer eBooks* (pp. 1–4). https://doi.org/10.1007/978-3-031-17125-3_343-1

[2] Kotseruba, I., & Tsotsos, J. K. (2018). 40 years of cognitive architectures: core cognitive abilities and practical applications.

[3] Langley, P., Laird, J. E., & Rogers, S. (2008). Cognitive architectures: Research issues and challenges.