

#35



Toward ethical cognitive architectures for the development of artificial moral agents. *Salvador Cervantes¹, Sonia López¹, José-Antonio Cervantes¹ (antonio.cervantes@valles.udg.mx).*

¹Department of Computing Science and Engineering, Universidad de Guadalajara, P.C. 46600, Ameca, Mexico.



Salvador Cervantes



Sonia López



José-Antonio Cervantes



ABSTRACT

Cognitive architectures aim to create some of the most challenging artificial agents commonly known as bio-inspired cognitive agents. This type of artificial agent seeks to embody human-like intelligence in order to operate and solve problems in the real world as humans do. Therefore, researchers in the machine ethics field face ethical questions related to what mechanisms an artificial agent must have for making moral decisions in order to ensure that their actions are always ethically right. This paper aims to identify some challenges that researchers need to solve in order to create ethical cognitive architectures. These cognitive architectures are characterized by the capacity to endow artificial agents with appropriate mechanisms to exhibit explicit ethical behavior.

INTRODUCTION

Every day, it is possible to find important applications where Artificial Agents (AAs) are becoming a major issue in today's digital society.



Figure 1. Self-driving car



Figure 2. Bio-inspired cognitive agents

COGNITIVE ARCHITECTURES

According to Kotseruba and Tsotsos (2018) the number of existing cognitive architectures has reached several hundred over the last 40 years. Some of these cognitive architectures, such as LIDA, Soar, ACT-R, and iCub try to be fundamental architectures for the Artificial General Intelligence (AGI) model of human cognition (Faghihi and Franklin, 2012; Kotseruba and Tsotsos, 2018; Lieto et al., 2018).

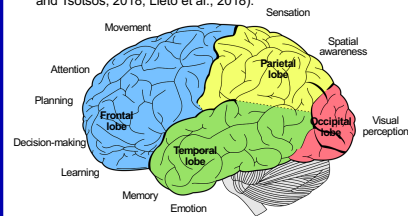


Figure 3. Cognitive architectures based on the human and non-human (primates) brains

A key consideration when developing this kind of bio-inspired cognitive agents is that they should be designed to be capable of coexisting in harmony with people and other systems.

MACHINE ETHICS

The issue of AAs and people coexisting in harmony has sparked a strong interest in the machine ethics field, which is a multi-disciplinary area that involves knowledge from computer science and moral philosophy. Researchers in this field deal with ethical questions related to what mechanisms AMAs must have for making moral decisions in order to ensure that their actions are always ethically right. According to Moor (2006, 2009), who is one of the pioneering theoreticians in the field of machine ethics, artificial ethical agents (commonly known as artificial moral agents) can be classified as follows:

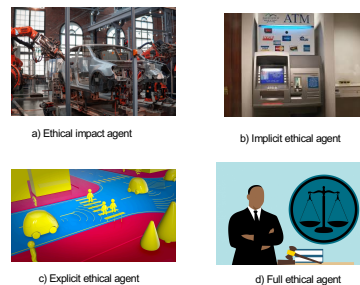


Figure 4. Taxonomy proposed by Moor

REASONS TO DEVELOP ETHICAL COGNITIVE ARCHITECTURES

This paper considers that one common error made in arguments to support the development of Artificial Moral Agents (AMAs) is the use of extreme outlier scenarios. These scenarios are philosophically interesting problems such as the trolley and tunnel problems (Cervantes et al., 2019; Gogoll and Müller, 2017), which are usually discussed in the literature under the rubric of "moral dilemmas." However, there are researchers who affirm that these examples of sacrificial dilemmas lack experimental, mundane, and psychological realism (Bauman et al., 2014; Kahane, 2015).

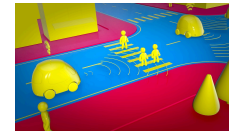


Figure 5. A variation of the trolley problem for self-driving cars

Ethical cognitive architectures are characterized by extending the cognitive functionality of AAs by endowing them with mechanisms to make ethical judgments. Two fundamental reasons that this paper considers for supporting the development of ethical cognitive architectures are:

- It contributes to a truly AGI model of human cognition. Some cognitive architectures want to be an AGI of human cognition (Faghihi and Franklin, 2012; Lieto et al., 2018). Therefore, developing a computational model of human beings' ethical decision-making is an essential cognitive function to achieve both a truly human-like intelligence in cognitive architectures and an AGI model of human cognition.



Figure 6. Evolution of artificial agents

- It extends the functionality and usage of AMAs. This reason is based on the hypothesis that AMAs developed by ethical cognitive architectures will be able to exhibit additional behaviors, in contrast with AAs developed by current cognitive architectures. For instance, cognitive architectures have been used to develop artificial social agents to study different aspects related to human social behavior (Wykowska et al., 2016). However, these AAs have limited use to study aspects related to human ethical behavior, because they are unable to process ethical information. Thus, ethical cognitive architectures may be suitable for use in such a situation and also in new domains where ethical and unethical behaviors are studied from different psychological and social approaches.



Figure 7. Artificial social agents

CHALLENGES FOR DEVELOPING ETHICAL COGNITIVE ARCHITECTURES

We address technological and social questions such as what moral cognitive functions an ethical cognitive architecture should have and what social conditions should exist to use AMAs in the real-world.

Moral cognitive functions needed to improve AMAs' behavior.

- Moral emotions.
- Moral agency.
- Autonomy.

Technical and social challenges that come with deploying AMAs in the real-world.

- Trusted ethical cognitive architectures.
- Standards and certifications.
- Legal framework.

CONCLUSIONS

Cognitive architectures aim to create AAs capable of exhibiting human-like behaviors in order to operate and solve problems in the real-world as humans do. The deployment of AAs in our society in the not-so-distant future is already being envisioned for many different applications, ranging from self-driving cars to robots for military purposes. This paper proposes the creation of ethical cognitive architectures in order to move toward AMAs capable of making ethical judgments.

This paper concludes that because of the current interaction between AAs and humans, ethical mechanisms for AAs are necessary in order to deal with the ethical, legal and societal impacts of the new generation of AAs. However, from technological, cultural, and social perspectives, there are still open challenges that must be faced before the new generation of autonomous systems based on AAs can operate in our society.

REFERENCES

- Bauman, C.W., McGraw, A.P., Bartels, D.M., Warren, C., 2014. Revisiting external validity: Concerns about trolley problems and other sacrificial dilemmas in moral psychology. *Social and Personality Psychology Compass* 8, 536-554. DOI:10.1111/spc3.12131.
- Cervantes, J.A., Lopez, S., Rodriguez, L.F., Cervantes, S., Cervantes, F., Ramos, F., 2019. Artificial moral agents: A survey of the current status. *Science and engineering ethics* , 1-32. DOI:10.1007/s11948-019-00151-x.
- Faghihi, U., Franklin, S., 2012. The LIDA model as a foundational architecture for AGI, in: *Theoretical Foundations of Artificial General Intelligence*, pp. 103-121. DOI:10.2991/978-94-91216-62-6_7.
- Gogoll, J., Müller, J.F., 2017. Autonomous cars: In favor of a mandatory ethics setting. *Science and engineering ethics* 23, 681-700. DOI:10.1007/s11948-016-9606-x.
- Kahane, G., 2015. Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Social neuroscience* 10, 551-560. DOI:10.1080/17470919.2015.1023400.
- Kotseruba, I., Tsotsos, J.K., 2018. 40 years of cognitive architectures: Core cognitive abilities and practical applications. *Artificial Intelligence Review*, 1-78. DOI:10.1007/s10462-018-9646-y.
- Lieto, A., Bhatt, M., Oltramari, A., Vernon, D., 2018. The role of cognitive architectures in general artificial intelligence. *Cognitive Systems Research* 48, 1-3. DOI:10.1016/j.cogsys.2017.08.003.
- Moor, J.H., 2009. Four kinds of ethical robots. *Philosophy Now* 72, 12-14.
- Moor, J.H., 2006. The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems* 21, 18-21. DOI:10.1109/IMS.2006.80.
- Wykowska, A., Chaminade, T., Cheng, G., 2016. Embodied artificial agents for understanding human social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences* 371, 20150375. DOI:10.1098/rstb.2015.0375.