Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Available online at www.sciencedirect.com

ScienceDirect

Cognitive Systems Research 64 (2020) 117–125

Cognitive Systems RESEARCH

www.elsevier.com/locate/cogsys

Toward ethical cognitive architectures for the development of artificial moral agents

Salvador Cervantes, Sonia López, José-Antonio Cervantes *

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

Received 15 June 2020; accepted 26 August 2020
Available online 3 September 2020

Abstract

New technologies based on artificial agents promise to change the next generation of autonomous systems and therefore our interaction with them. Systems based on artificial agents such as self-driving cars and social robots are examples of this technology that is seeking to improve the quality of people's life. 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. Moreover, some cognitive architectures such as Soar, LIDA, ACT-R, and iCub try to be fundamental architectures for the Artificial General Intelligence model of human cognition. 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. Additionally, we offer some reasons to develop ethical cognitive architectures. We hope that this study can be useful to guide future research on ethical cognitive architectures.

Keywords: Ethical cognitive architectures; Cognitive functions; Artificial agents; Machine ethics; Artificial moral agents

1. Introduction

Our inventions have constantly redefined our style of life. For instance, they have changed our way of working, communicating, and traveling. In recent decades, technology has become highly embedded in people's daily life. Advances in the Artificial Intelligence field have allowed humans to seek different alternatives to delegate part of their decision-making power to Artificial Agents (AAs) (Cervantes et al., 2019; Ingrand & Ghallab, 2017; Mostafa, Ahmad, & Mustapha, 2017). These AAs are characterized by their capabilities for perceiving and interacting with the environment. Every day, it is possible to find important applications where AAs are becoming a major issue in today's digital society (Cervantes et al., 2019; Gutierrez-Garcia & Rodríguez, 2016; Mostafa et al., 2017). Two relevant examples of these applications are: (1) self-driving cars, whose benefits in terms of safety, efficiency, environmental impacts, and increased mobility are well-known (Haboucha, Ishaq, & Shiftan, 2017) and (2) bio-inspired cognitive agents that have been used for developing a wide variety of practical applications encompassing games and puzzles, robotics, psychological experiments, natural language processing, human-robot and human–computer interaction, computer vision, and virtual

* Corresponding author.
E-mail address: antoniocervantes@valles.udg.mx (J.-A. Cervantes).

https://doi.org/10.1016/j.cogsys.2020.08.010
1389-0417/

agents (Kotseruba & Tsotsos, 2018). It is remarkable that AAs are endowed with a set of complex functions such as perception, memory, planning, decision-making, and in some cases, these AAs include mechanisms for improving their behavior based on previous experiences (Barbosa, Leitão, Adam, & Trentesaux, 2015; Romero, 2019).

The computing research community has made efforts over the last few decades to create AAs capable of embodying human-like intelligence (Franklin, Madl, D'mello, & Snaider, 2013; Kotseruba & Tsotsos, 2018). 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 (Faghihi & Franklin, 2012; Franklin et al., 2013), Soar (Laird, 2008), ACT-R (Borst & Anderson, 2015), and iCub (Metta et al., 2010; Natale et al., 2013) try to be fundamental architectures for the Artificial General Intelligence (AGI) model of human cognition (Faghihi & Franklin, 2012; Kotseruba & Tsotsos, 2018; Lieto, Bhatt, Oltramari, & Vernon, 2018). These cognitive architectures seek to create both physical and virtual AAs capable of exhibiting intelligent behavior in a general setting, the way human agents do. For instance, these AAs must be able to adapt and learn how to behave in new situations and invent new solutions on the basis of past experiences as human agents do (Lieto et al., 2018; Metta et al., 2010). 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. This issue of coexistence and interaction between AAs and people has given rise to a new study field known as machine ethics (Cervantes et al., 2019). Researchers in machine ethics have proposed endowing AAs with appropriate mechanisms in order to give them the ability to deal with ethical problems. Most approaches for developing these mechanisms have taken inspiration from normative ethical theories such as teleological and deontological theories (Cervantes et al., 2019; Goodall, 2014). However, from the academic domain, a variety of scholars in the fields of ethics, technology, and machine ethics are debating whether creating ethical machines is a right or wrong approach (Arkin, 2018; Etzioni & Etzioni, 2017; Malle, 2016; Wykowska, Chaminade, & Cheng, 2016). Rather than contributing to this debate, the objective of the present paper is to analyze some challenges for developing ethical cognitive architectures. These cognitive architectures are characterized by the capacity to endow AAs with mechanisms to exhibit (explicit) ethical behavior. Such AAs are commonly known as Artificial Moral Agents (AMAs). Additionally, we present some reasons to support the development of ethical cognitive architectures. The remainder of this paper is structured as follows. In the "Cognitive architectures and ethical agents" section, a detailed explanation of Artificial Moral Agents is presented, and an argument is made that it is important to develop ethical cognitive architectures. In the "Challenges for developing ethical cognitive architectures" section, technological and social challenges for creating ethical cognitive architectures are described. Finally, the "Conclusion" section provides some concluding remarks about the work done so far to develop ethical cognitive architectures.

2. Cognitive architectures and ethical agents

The term "*Cognitive Architectures*" indicates abstract models of cognition in natural and artificial agents (Lieto et al., 2018). Cognitive architectures are a part of research in AGI, which is seeking to create AAs capable of embodying human-like intelligence. Most AAs reported in the literature can be grouped into two categories: (1) AAs that act like human beings (Franklin et al., 2013; Gutierrez-Garcia & Rodríguez, 2016; Metta et al., 2010) and (2) AAs that act rationally (Barbosa et al., 2015; Omohundro, 2012). AAs that act like human agents have the chance to make errors, but such errors must be like the errors typically made by human agents in similar situations, in contrast to AAs that act rationally, which are required to produce consistent and correct behaviors for arbitrary tasks (e.g. a chess robot).

Independently of whether an AA is able to act like human agents or rationally, the rapid development of AAs promises to change the next generation of autonomous systems and therefore our interaction with them. Smart cities (Batty et al., 2012) and cognitive assisted living ambient systems (Li, Lu, & McDonald-Maier, 2015) are examples of environments that could be governed by AAs. These smart and cognitive environments promise to improve the quality of people's life. For instance, people with special needs such as older adults or disabled people will be able to live independently and comfortably as long as possible in their living environment. People's living environments are not limited to their home, but also encompass various environments such as neighborhoods, shopping malls, and other public places. Applications of smart cities and cognitive assisted living ambient systems consist of complex networks of heterogeneous information appliances and smart artifacts such as self-driving and cognitive vehicles (Haboucha et al., 2017; Zhang, Zhou, Liu, & Hussain, 2019), smart wheelchairs (Schwesinger, Shariati, Montella, & Spletzer, 2017), smart homes (Feng, Setoodeh, & Haykin, 2017), social and cognitive robots (Leite, Martinho, & Paiva, 2013; Metta et al., 2010), among other artifacts that can assist people in a variety of ways.

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 (Cervantes et al., 2019). 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), who is one of the pioneering theoreticians in the field of machine ethics, artificial ethical agents can be classified as follows:

- *Ethical impact agents.* These AAs would seem to be *"ethical agents"* in the weakest sense, being those AAs whose actions have an (indirect) ethical impact, whether intended or not. Any physical or virtual agent is a potential ethical impact agent to the extent that its actions could cause harm or benefit to humans (Moor, 2009). For instance, there are robots designed to protect people from danger by performing tasks that are harmful to people's health. There are also AAs designed to enhance the quality of people's lives by performing tedious jobs.
- *Implicit ethical agents.* These AAs are unable to distinguish between ethical and unethical behaviors; however, their design involves safety or critical reliability concerns to avoid unethical behaviors (Moor, 2006). For instance, banking transactions are ethically important because they involve money. In this context, banks' automated teller machines are implicit ethical agents because they do not need to know what actions are ethically right or not. Automated teller machines only need to be carefully constructed to give out or transfer the correct amount of money every time a banking transaction occurs.
- *Explicit ethical agents.* These are AAs that can identify and process ethical information about a variety of situations and make explicit ethical judgments (Moor, 2006). These AAs are commonly known as AMAs to indicate that they have been programed with ethical mechanisms or functions to make explicit ethical decisions (Cervantes et al., 2019; Wallach, Allen, & Smit, 2008). Most strategies reported in the literature to create these AAs take their inspiration from normative ethical theories such as utilitarian and deontological theories (Cervantes et al., 2019; Goodall, 2014).
- *Full ethical agents.* Like explicit ethical agents, full ethical agents make ethical judgments about a wide variety of situations. However, full ethical agents have those central metaphysical features (such as *"consciousness, intentionality, and free will"*) that we usually attribute to ethical agents like human beings. Therefore, only human beings are considered capable of being fully ethical (Moor, 2009).

We are aware that many machine ethics researchers have offered reasons to support the development of explicit ethical agents or AMAs. Some of those reasons have been critiqued by other researchers such as Etzioni and Etzioni (2017) and van Wynsberghe and Robbins (2019). This paper considers that one common error made in arguments to support the development of 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 & 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, McGraw, Bartels, & Warren, 2014; Kahane, 2015). Furthermore, these sacrificial dilemmas are unrealistic and unrepresentative of the moral situations people encounter in their everyday life. Nevertheless, this paper considers that studies done in the machine ethics field are essential for developing both ethical cognitive architectures and AMAs. 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 (1) it contributes to a truly AGI model of human cognition, and (2) it extends the functionality and usage of AMAs. The first reason is based on the fact that cognitive architectures seek to create AAs capable of embodying human-like intelligence. Furthermore, some cognitive architectures want to be an AGI of human cognition (Faghihi & 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. The second 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. For instance, social behavior researchers can use ethical cognitive architectures to create agent-based virtual simulations of social systems susceptible to corruption in order to study causes, effects and how to address different corruption issues (Gutierrez-Garcia & Rodríguez, 2016). Another interesting example is the situation experienced by people around the world as a consequence of pandemic diseases such as Spanish flu, H1N1 influenza, and COVID-19. An agent-based virtual simulation system could be useful to study people's ethical behavior in dangerous and stressful situations like those generated by pandemic diseases, where people must commonly follow strict security norms imposed by their governments to deal with such situations. Moreover, benefits offered by robots that help in the health area have been studied and analyzed from several decades (Bemelmans, Gelderblom, Jonker, & De Witte, 2012; Greczek, Kaszubski, Atrash, & Matarić, 2014; Qureshi & Syed, 2014). However, many relevant ethical concerns have arisen, such as What possible benefits/limitations could be associated with robotics with respect to employment and employees? Could caregiving robots displace and/or de-skill human caregivers? Could a robot provide good care? And is there an appropriate normative framework

to regulate the use of these robots? These examples aim to show how ethical cognitive architectures can extend the functionality of AAs. Additionally, AMAs try to be a feasible computational approach to deal with moral situations that may arise in interaction with humans (Cervantes et al., 2019).

3. Challenges for developing ethical cognitive architectures

Machine ethics encompasses questions about what moral cognitive functions AMAs should have and how these cognitive functions could be computationally implemented. As commented before, this paper considers that an ethical cognitive architecture is a cognitive architecture that can create AMAs capable of making ethical judgments. Therefore, in this section, 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. The following points analyze some moral cognitive functions needed to improve AMAs' behavior. Designing computational models of these moral cognitive functions involves technological challenges that researchers and engineers need to solve to create ethical cognitive architectures.

- *Moral emotions.* Ethical cognitive architectures need to endow AMAs with a set of ethical functions if the AMAs are to exhibit ethical behavior. These ethical functions commonly include an explicit set of well-defined and coherent ethical rules and norms designed to produce consistent ethical judgments (Cervantes et al., 2019). However, AMAs are not exempt from facing moral conflicts as a result of interacting with other artificial entities or human beings. This is because the world is an open and dynamic place where unexpected events or situations may arise. Additionally, AMAs and people do not necessarily share the same beliefs and moral values. An alternative to deal with such conflicts is to endow AMAs with moral emotions. Therefore, ethical cognitive architectures need to consider some underlying processes related to emotions and their regulation in order to implement moral emotions in AMAs. Studies reported in the literature show how moral emotions play an important role in determining how human beings deal with moral conflict situations (Krettenauer, Colasante, Buchmann, & Malti, 2014; Malti & Latzko, 2010). Some AMAs and cognitive architectures reported in the literature have proposed algorithms inspired by the human emotional system to offer a mechanism capable of using anticipated moral emotions as a moral regulator of the ethical decision-making process, dispensing favorable or unfavorable (punishment) rewards from an ethical approach (Cervantes, Rodríguez, López, Ramos, & Robles, 2016; Wallach, Franklin, & Allen, 2010). A system of moral emotions can be useful to implement self-evaluative moral emotions, which arise either ex-ante or ex-post an ethical or unethical action (Krettenauer et al., 2014). In other words, a system of moral emotions offers AMAs the possibility to compute both positive and negative moral emotions such as pride and guilt in anticipation of respecting or violating a moral norm or after the occurrence of an ethical or unethical action (Krettenauer et al., 2014). Thus, moral emotions reflect the self-relevance of moral rules and values. This process could provide critical information about the desirability of future actions and can thus be seen as an important predictor of moral decision-making (Cervantes et al., 2016). Machine ethics researchers suggest that incremental knowledge about emotional and affective human processes may lead to the development of human-like moral agency in AMAs, or something close to it (Wallach et al., 2010).
- *Moral agency.* This is a key point for developing ethical cognitive architectures capable of endowing AMAs with appropriate mechanisms for coexisting in harmony with people and other systems. Agency is the property of agents that allows them to communicate and negotiate with other agents (human or not) based on different principles, in order to carry out cooperative tasks. Thus, an AMA incapable of communicating and cooperating with other AMAs is a limited system for coexisting in an open, dynamic, and heterogeneous environment. For instance, sometimes an AMA needs to deal with situations where its behavior can be influenced by external situational factors involving other AMAs. It might seem extremely difficult or perhaps impossible for AMAs based on different norms and ethical theories to cooperate to solve such situations. In fact, this task can be difficult to achieve even when AMAs are based on the same ethical theory but different underlying criteria. Mechanisms related to empathy are basic for endowing AMAs with human-like moral agency, because empathy is the capacity to enter sympathetically into the concerns and feelings of others in order to think in terms of their interests (Aaltola, 2014). Also, an empathy system is built on certain cognitive functions such as social cognition, mental state, and reasoning. These are basic fundamental functions for developing an empathy system. Social cognition is the ability to think about the contents of someone else's mind in order to directly link first and third-agent experiences (Gallese, Keysers, & Rizzolatti, 2004). Mental state is the ability of an AMA to step into the other's shoes, and simulate different mental states in the circumstances it faces (Aaltola, 2014). This cognitive function calls for AMAs that can reflect upon their own and others' beliefs, knowledge, and intentions. Finally, reasoning allows AMAs to perform conscious deliberations to reach some consistency among their competing desires and values in order to cooperate with others, while acting as consistently as possible according to their own principles (Nahmias, 2007). Developing computational models of these functions to endow AMAs with

a truly moral agency represents a formidable challenge for researchers seeking to create ethical cognitive architectures.

- *Autonomy.* There are several definitions of the concept of autonomy. One of the oldest of these comes from ancient Greece and refers to *"one who gives oneself their own law"* (Thórisson & Helgasson, 2012). In modern language the line between autonomy and automation has a tendency to get blurred. The term automation is used to refer to a system's ability to perform a task without the intervention of an external operator (Thórisson & Helgasson, 2012). A washing machine is an automatic machine, which after being started by a human operator is able to change from one cycle to another, taking the clothes right through from start to finish without any additional input from a human operator. In the same way, a self-driving car capable of driving itself from one city to another without a human operator might be viewed as an *"automatic car"* rather than *"autonomous car"*. Thus, autonomy seems to require something above and beyond even fairly sophisticated automation. However, this issue is still open for debate (Abbass, Petraki, Merrick, Harvey, & Barlow, 2016). According to Abbass et al. (2016) autonomy comprises aspects of self-governance and suggests that agents rely on their own laws and work independently. To achieve true autonomy in AMAs, ethical cognitive architectures must endow AMAs with a set of norms and laws that govern their decision-making, but also, a set of cognitive functions for learning and reasoning that allows AMAs to acquire new knowledge, behaviors, skills, values, or preferences, or to modify the ones they already have, in order to improve their behavior. Moreover, in order to achieve self-governance in AMAs, it is essential to endow them with a motivational system. Motivation has a key role in driving and regulating cognitive and social behavior in both humans and AAs (Bach, 2011; Cosmides & Tooby, 2013). Psychological theories described in the literature show how motivation pushes or influences human beings to fulfill wants or needs, such as physiological needs, safety needs, love and belonging, esteem, and self-actualization (Taormina & Gao, 2013).

In addition to challenges associated with developing computational models of the moral cognitive functions described above, there are technical and social challenges that come with deploying AMAs in the real world. Some of these challenges are:

- *Trusted ethical cognitive architectures.* Trust is a key point for encouraging the adoption of any new technology by human beings. Thus, AMA adoption is a term that refers to the acceptance, integration, and use of AMAs in society. Getting people to trust AMAs is not a trivial matter, and it becomes even more complex as we move from AAs to AMAs. Explainable Artificial Intelligence and formal proofs for moral reasoning are two different approaches used to achieve trust in both AAs and AMAs. The notion of explainable Artificial Intelligence has been studied for the past three decades. However, it has drawn increasing interest because many Artificial Intelligence applications have limited take-up, or are not appropriated at all, due to ethical concerns and a lack of trust on the part of their users (Miller, 2019). On the other hand, formal proofs for moral reasoning is a relatively new area of study in machine ethics. Nowadays, there are few frameworks or methods reported in the literature to formally evaluate ethical rules implemented in AMAs (Cervantes et al., 2019). These frameworks aim to offer a formal specification and verification of both AMAs' ethical reasoning and their behavior in order to achieve trusted AMAs. However, there are a number of aspects that researchers and engineers need to take into account in order to achieve people's trust in this technology. These aspects include standards, responsible research and innovation, public engagement, as well as the generation of laws for regulating the right use of AMAs in society.
- *Standards and certifications.* Standards are international agreements by experts regarding the best practices for making a product, managing a process, delivering a service or supplying materials. Thus, standards are a vital part of the software development process. All standards embody an ethical principle or value because they are a guide for developing quality software. Therefore, all standards can be thought of as implicit ethical standards. However, developers in the machine ethics field need explicit ethical standards that address clearly articulated ethical concerns for developing and testing ethical cognitive architectures and AMAs. The use of standards is the first step to define quality certification mechanisms for these systems. Creating standards for guiding the development of ethical cognitive architectures for AMAs is a complex challenge that researchers must take into account. Currently, expert groups focused on Artificial Intelligence around the world are working to define guidelines and standards for Ethical Artificial Intelligence (Hagendorff, 2020; Jobin, Ienca, & Vayena, 2019). The greatest effort has been made the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. The IEEE Global Initiative is working on inspiring the creation of standards and associated certification programs (IEEE, 2017a). As a result, a new generation of ethical standards in robotics and Artificial Intelligence is emerging. These standards are:
  - *IEEE P7000 Model Process for Addressing Ethical Concerns During System Design.* The IEEE P7000 standard establishes a process model by which engineers can address ethical consideration throughout various stages of system initiation, analysis, and design (IEEE, 2016a).

- *IEEE P7001 Transparency of Autonomous Systems.* This standard describes measurable, testable levels of transparency, so that autonomous systems can be objectively assessed and levels of compliance determined (IEEE, 2016b).
- *IEEE P7002 Data Privacy Process.* The IEEE P7002 standard defines privacy requirements for software engineering processes that develop systems utilizing employee, customer or other external users' personal data (IEEE, 2016c).
- *IEEE P7003 Algorithmic Bias Considerations.* This standard describes specific methodologies to help users certify how they worked to address and eliminate issues of negative bias in the creation of their algorithms. Negative bias infers the usage of overly subjective criteria or information known to be inconsistent with legislation concerning certain protected characteristics such as race, gender, and sexuality (IEEE, 2017b).
- *IEEE P7004 Standard on Child and Student Data Governance.* The IEEE P7004 standard defines specific methodologies to help users certify how they approach accessing, collecting, storing, utilizing, sharing, and destroying child and student data. The standard provides specific metrics and compliance criteria regarding these types of uses from trusted global partners and specifies how vendors and educational institutions can meet them (IEEE, 2017c).
- *IEEE P7005 Standard for Transparent Employer Data Governance.* This standard defines specific methodologies to help employers to certify how they approach accessing, collecting, storing, utilizing, sharing, and destroying employee data. The standard provides specific metrics and specifies criteria regarding these types of uses from trusted global partners and specifies how vendors and employers can meet them (IEEE, 2017d).
- *IEEE P7006 Standard for Personal Data Artificial Intelligence (AI) Agent.* The IEEE P7006 standard describes the technical elements required to create and grant access to a personalized AA that will comprise inputs, learning, ethics, rules and values controlled by people (IEEE, 2017e).
- *IEEE P7007 Ontological Standard for Ethically Driven Robotics and Automation Systems.* This standard establishes a set of ontologies with different abstraction levels that contain concepts, definitions and axioms which are necessary to establish ethically driven methodologies for the design of Robots and Automation Systems (IEEE, 2017f).
- *IEEE P7008 Standard for Ethically Driven Nudging for Robotic, Intelligent, and Automation Systems.* The IEEE P7008 standard establishes a delineation of typical nudges (nudges are defined as overt or hidden suggestions or manipulations designed to influence the behavior or emotions of a user). It contains concepts, functions and benefits necessary to establish and ensure ethically driven methodologies for the design of the robotic, intelligent and autonomous systems that incorporate them (IEEE, 2017g).
- *IEEE P7009 Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems.* This standard establishes a practical, technical baseline of specific methodologies and tools for the development, implementation, and use of effective fail-safe mechanisms in autonomous and semi-autonomous systems (IEEE, 2017h).
- *IEEE P7010 Well-being Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems.* This standard establishes mechanisms for measuring the impact of Artificial Intelligence or autonomous and intelligent systems on humans' well-being (IEEE, 2020). Additionally, The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (IEEE, 2017a) has proposed five general principles to guide the ethical design, development, and implementation of autonomous and semi-autonomous AAs (including AMAs):
- *Human Rights.* Ensure they do not infringe on internationally recognized human rights.
- *Well-being.* Prioritize metrics of well-being in their design and use.
- *Accountability.* Ensure that their designers and operators are responsible and accountable.
- *Transparency.* Ensure they operate in a transparent manner.
- *Awareness of misuse.* Minimize the risks of their misuse.

- *Legal framework.* The lack of a legal framework for using full autonomous AMAs in real environments has given rise to legal questions such as who should be responsible when an AMA harms something or someone (the owner of the AMA, its developer or the person who sells it). For instance, a self-driving car is not exempt from being involved in traffic accidents. In such a situation, who should be held legally and morally responsible for remedying the harm caused by a vehicle that carries only a passenger (who is the owner of the self-driving car), but no driver? Researchers have analyzed this challenge from different viewpoints, such as AMAs' design and their legal implications (De Sio, 2017; Gogoll & Müller, 2017; Schreurs & Steuwer, 2015; Taeihagh & Lim, 2019). From a design approach, different alternatives have been proposed in the literature such as mandatory ethics setting (Gogoll & Müller, 2017), personal ethics setting (Contissa, Lagioia, & Sartor, 2017), and adjustable autonomy (Mostafa et al., 2017). However, from a legal approach, there is limited or no country-specific literature regarding the normative implications of AMAs such as self-driving cars (Taeihagh & Lim, 2019) and robots in healthcare (Stahl & Coeckelbergh, 2016). Philosophical

studies on the regulation of self-driving cars show that governments have in most instances avoided stringent measures in order to promote self-driving car developments (Schreurs & Steuwer, 2015; Taeihagh & Lim, 2019). The United States was the first government worldwide to provide licenses for the testing and operation of self-driving cars, albeit under strict conditions (Schreurs & Steuwer, 2015). However, regulatory gaps have been identified in governments' approach to safety risks involving self-driving cars (Taeihagh & Lim, 2019). A similar situation is presented in other domains such as robots in healthcare (Stahl & Coeckelbergh, 2016) and robots for military purposes (Arkin, 2018), where there is no appropriate legal framework to regulate such technology. The current discourse on legal risks and responsibilities of both AAs and AMAs reported in the literature shows a rich landscape of inquiry in the area; it is important to understand and address the challenges that governments face in defining a full legal framework for regulating the use of both AAs and AMAs in different domains.

4. Conclusion

As described in this paper, 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. Moreover, human beings are looking to delegate part of their decision-making power to AAs, thus increasing the scope of their activities. 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. Endowing AAs with ethical mechanisms has been an initial approach in the field of machine ethics that attempts to deal with some issues likely to arise in the relationship between AAs and human beings so that they may coexist in a safe way, in harmony, and with confidence. This paper presented some reasons to develop ethical cognitive architectures capable of endowing AAs with ethical mechanisms. As a result of this study, some challenges that researchers need to face for developing ethical cognitive architectures were identified and analyzed. Additionally, a new generation of ethical standards in robotics and Artificial Intelligence was identified in the literature. 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. We hope that this study can be useful to guide future research on ethical cognitive architectures.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Aaltola, E. (2014). Varieties of empathy and moral agency. *Topoi, 33*(1), 243–253.

Abbass, H. A., Petraki, E., Merrick, K., Harvey, J., & Barlow, M. (2016). Trusted autonomy and cognitive cyber symbiosis: Open challenges. *Cognitive Computation, 8*(3), 385–408.

Arkin, R. (2018). Lethal autonomous systems and the plight of the non-combatant. In The political economy of robots (pp. 317–326).

Bach, J. (2011). A motivational system for cognitive ai. In *International conference on artificial general intelligence* (pp. 232–242).

Barbosa, J., Leitão, P., Adam, E., & Trentesaux, D. (2015). Dynamic self-organization in holonic multi-agent manufacturing systems: The ADACOR evolution. *Computers in Industry, 66*, 99–111.

Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., Ouzounis, G., & Portugali, Y. (2012). Smart cities of the future. *The European Physical Journal Special Topics, 214*(1), 481–518.

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*(9), 536–554.

Bemelmans, R., Gelderblom, G. J., Jonker, P., & De Witte, L. (2012). Socially assistive robots in elderly care: A systematic review into effects and effectiveness. *Journal of the American Medical Directors Association, 13*(2), 114–120.

Borst, J. P., Anderson, J. R. (2015). Using the ACT-R cognitive architecture in combination with fMRI data. In An introduction to model-based cognitive neuroscience (pp. 339–352).

Cervantes, J.-A., López, S., Rodríguez, L.-F., Cervantes, S., Cervantes, F., & Ramos, F. (2019). Artificial moral agents: A survey of the current status. *Science and Engineering Ethics*, 1–32.

Cervantes, J.-A., Rodríguez, L.-F., López, S., Ramos, F., & Robles, F. (2016). Autonomous agents and ethical decision-making. *Cognitive Computation, 8*(2), 278–296.

Contissa, G., Lagioia, F., & Sartor, G. (2017). The ethical knob: ethically-customisable automated vehicles and the law. *Artificial Intelligence and Law, 25*(3), 365–378.

Cosmides, L., & Tooby, J. (2013). Evolutionary psychology: New perspectives on cognition and motivation. *Annual Review of Psychology, 64*, 201–229.

De Sio, F. S. (2017). Killing by autonomous vehicles and the legal doctrine of necessity. *Ethical Theory and Moral Practice, 20*(2), 411–429.

Etzioni, A., & Etzioni, O. (2017). Incorporating ethics into artificial intelligence. *The Journal of Ethics, 21*(4), 403–418.

Faghihi, U., Franklin, S. (2012). The LIDA model as a foundational architecture for AGI. In Theoretical Foundations of Artificial General Intelligence (pp. 103–121).

Feng, S., Setoodeh, P., & Haykin, S. (2017). Smart home: Cognitive interactive people-centric internet of things. *IEEE Communications Magazine, 55*(2), 34–39.

Franklin, S., Madl, T., D'mello, S., & Snaider, J. (2013). LIDA: A systems-level architecture for cognition, emotion, and learning. *IEEE Transactions on Autonomous Mental Development, 6*(1), 19–41.

Gallese, V., Keysers, C., & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences, 8*(9), 396–403.

Gogoll, J., & Müller, J. F. (2017). Autonomous cars: In favor of a mandatory ethics setting. *Science and Engineering Ethics, 23*(3), 681–700.

Goodall, N. J. (2014). Machine ethics and automated vehicles. In Road vehicle automation (pp. 93–102).

Greczek, J., Kaszubski, E., Atrash, A., & Matarić, M. (2014). Graded cueing feedback in robot-mediated imitation practice for children with autism spectrum disorders. In *The 23rd IEEE international symposium on robot and human interactive communication* (pp. 561–566).

Gutierrez-Garcia, J. O., & Rodríguez, L.-F. (2016). Corruptible social agents. *Computer Animation and Virtual Worlds, 27*(2), 89–102.

Haboucha, C. J., Ishaq, R., & Shiftan, Y. (2017). User preferences regarding autonomous vehicles. *Transportation Research Part C: Emerging Technologies, 78*, 37–49.

Hagendorff, T. (2020). The ethics of ai ethics: An evaluation of guidelines. *Minds and Machines*, 1–22.

IEEE, 2016a. P7000 - model process for addressing ethical concerns during system design. Accessed 10 April 2020. URL <https://standards.ieee.org/project/7000.html>.

IEEE, 2016b. P7001 - transparency of autonomous systems. Accessed 10 April 2020. URL <https://standards.ieee.org/project/7001.html>.

IEEE, 2016c. P7002 - data privacy process. Accessed 10 April 2020. URL <https://standards.ieee.org/project/7002.html>.

IEEE, 2017a. Aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems, version 2. Tech. rep., IEEE, accessed 10 April 2020. URL <https://standards.ieee.org/industry-connections/ec/ead-v1.html>.

IEEE, 2017b. P7003 - algorithmic bias considerations. Accessed 11 April 2020. URL <https://standards.ieee.org/project/7003.html>.

IEEE, 2017c. P7004 - standard for child and student data governance. Accessed 11 April 2020. URL <https://standards.ieee.org/project/7004.html>.

IEEE, 2017d. P7005 - standard for transparent employer data governance. Accessed 11 April 2020. URL <https://standards.ieee.org/project/7005.html>.

IEEE, 2017e. P7006 - standard for personal data artificial intelligence (ai) agent. Accessed 11 April 2020. URL <https://standards.ieee.org/project/7006.html>.

IEEE, 2017f. P7007 - ontological standard for ethically driven robotics and automation systems. Accessed 12 April 2020. URL <https://standards.ieee.org/project/7007.html>.

IEEE, 2017g. P7008 - standard for ethically driven nudging for robotic, intelligent and autonomous systems. Accessed 12 April 2020. URL <https://standards.ieee.org/project/7008.html>.

IEEE, 2017h. P7009 - standard for fail-safe design of autonomous and semi-autonomous systems. Accessed 12 April 2020. URL https://standards.ieee.org/project/7009.html.

IEEE, 2020. P7010 well-being metrics standard for ethical artificial intelligence and autonomous systems. Accessed 12 April 2020. URL https://standards.ieee.org/content/ieee-standards/en/standard/7010-2020.html.

Ingrand, F., & Ghallab, M. (2017). Deliberation for autonomous robots: A survey. *Artificial Intelligence, 247*, 10–44.

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of ai ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399.

Kahane, G. (2015). Sidetracked by trolleys: Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Social Neuroscience, 10*(5), 551–560.

Kotseruba, I., & Tsotsos, J. K. (2018). 40 years of cognitive architectures: Core cognitive abilities and practical applications. *Artificial Intelligence Review*, 1–78.

Krettenauer, T., Colasante, T., Buchmann, M., & Malti, T. (2014). The development of moral emotions and decision-making from adolescence to early adulthood: A 6-year longitudinal study. *Journal of Youth and Adolescence, 43*(4), 583–596.

Laird, J. E. (2008). Extending the soar cognitive architecture. *Frontiers in Artificial Intelligence and Applications, 171*, 224.

Leite, I., Martinho, C., & Paiva, A. (2013). Social robots for long-term interaction: A survey. *International Journal of Social Robotics, 5*(2), 291–308.

Li, R., Lu, B., & McDonald-Maier, K. D. (2015). Cognitive assisted living ambient system: A survey. *Digital Communications and Networks, 1*(4), 229–252.

Lieto, A., Bhatt, M., Oltramari, A., & Vernon, D. (2018). The role of cognitive architectures in general artificial intelligence. *Cognitive Systems Research, 48*, 1–3.

Malle, B. F. (2016). Integrating robot ethics and machine morality: The study and design of moral competence in robots. *Ethics and Information Technology, 18*(4), 243–256.

Malti, T., & Latzko, B. (2010). Children's moral emotions and moral cognition: Towards an integrative perspective. *New Directions for Child and Adolescent Development, 2010*(129), 1–10.

Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., Von Hofsten, C., Rosander, K., Lopes, M., Santos-Victor, J., et al. (2010). The iCub humanoid robot: An open-systems platform for research in cognitive development. *Neural Networks, 23*(8–9), 1125–1134.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38.

Moor, J. (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*(4), 18–21.

Mostafa, S. A., Ahmad, M. S., & Mustapha, A. (2017). Adjustable autonomy: A systematic literature review. *Artificial Intelligence Review, 51*(2), 149–186.

Nahmias, E. (2007). Autonomous agency and social psychology. In Cartographies of the Mind (pp. 169–185).

Natale, L., Nori, F., Metta, G., Fumagalli, M., Ivaldi, S., Pattacini, U., Randazzo, M., Schmitz, A., Sandini, G. (2013). The iCub platform: A tool for studying intrinsically motivated learning. In Intrinsically motivated learning in natural and artificial systems (pp. 433–458).

Omohundro, S. (2012). Rational artificial intelligence for the greater good. In Singularity Hypotheses (pp. 161–179).

Qureshi, M. O., & Syed, R. S. (2014). The impact of robotics on employment and motivation of employees in the service sector, with special reference to health care. *Safety and Health at Work, 5*(4), 198–202.

Romero, O. J. (2019). Cognitively-inspired agent-based service composition for mobile and pervasive computing. In *International Conference on AI and Mobile Services* (pp. 101–117).

Schreurs, M. A., Steuwer, S. D. (2015). Autonomous driving-political, legal, social, and sustainability dimensions. In Autonomes Fahren (pp. 151–173).

Schwesinger, D., Shariati, A., Montella, C., & Spletzer, J. (2017). A smart wheelchair ecosystem for autonomous navigation in urban environments. *Autonomous Robots, 41*(3), 519–538.

Stahl, B. C., & Coeckelbergh, M. (2016). Ethics of healthcare robotics: Towards responsible research and innovation. *Robotics and Autonomous Systems, 86*, 152–161.

Taeihagh, A., & Lim, H. S. M. (2019). Governing autonomous vehicles: Emerging responses for safety, liability, privacy, cybersecurity, and industry risks. *Transport Reviews, 39*(1), 103–128.

Taormina, R. J., & Gao, J. H. (2013). Maslow and the motivation hierarchy: Measuring satisfaction of the needs. *The American Journal of Psychology, 126*(2), 155–177.

Thórisson, K., & Helgasson, H. (2012). Cognitive architectures and autonomy: A comparative review. *Journal of Artificial General Intelligence, 3*(2), 1–30.

van Wynsberghe, A., & Robbins, S. (2019). Critiquing the reasons for making artificial moral agents. *Science and Engineering Ethics, 25*(3), 719–735.

Wallach, W., Allen, C., & Smit, I. (2008). Machine morality: Bottom-up and top-down approaches for modelling human moral faculties. *Ai & Society, 22*(4), 565–582.

Wallach, W., Franklin, S., & Allen, C. (2010). A conceptual and computational model of moral decision making in human and artificial agents. *Topics in Cognitive Science, 2*(3), 454–485.

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*(1693), 20150375.

Zhang, X., Zhou, M., Liu, H., & Hussain, A. (2019). A cognitively inspired system architecture for the mengshi cognitive vehicle. *Cognitive Computation*, 1–10.