Science and Engineering Ethics (2020) 26:501–532
https://doi.org/10.1007/s11948-019-00151-x

REVIEW

Artificial Moral Agents: A Survey of the Current Status

José-Antonio Cervantes[1] · Sonia López[1] · Luis-Felipe Rodríguez[2] · Salvador Cervantes[1] · Francisco Cervantes[3] · Félix Ramos[4]

Received: 12 July 2018 / Accepted: 17 October 2019 / Published online: 12 November 2019

Abstract

One of the objectives in the field of artificial intelligence for some decades has been the development of artificial agents capable of coexisting in harmony with people and other systems. The computing research community has made efforts to design artificial agents capable of doing tasks the way people do, tasks requiring cognitive mechanisms such as planning, decision-making, and learning. The application domains of such software agents are evident nowadays. Humans are experiencing the inclusion of artificial agents in their environment as unmanned vehicles, intelligent houses, and humanoid robots capable of caring for people. In this context, research in the field of machine ethics has become more than a hot topic. Machine ethics focuses on developing ethical mechanisms for artificial agents to be capable of engaging in moral behavior. However, there are still crucial challenges in the development of truly *Artificial Moral Agents*. This paper aims to show the current status of Artificial Moral Agents by analyzing models proposed over the past two decades. As a result of this review, a taxonomy to classify Artificial Moral Agents according to the strategies and criteria used to deal with ethical problems is proposed. The presented review aims to illustrate (1) the complexity of designing and developing ethical mechanisms for this type of agent, and (2) that there is a long way to go (from a technological perspective) before this type of artificial agent can replace human judgment in difficult, surprising or ambiguous moral situations.

Keywords Artificial agent · Ethical agent · Moral dilemma · Machine ethics

Introduction

An objective in the field of artificial intelligence (AI) for some decades has been the development of Artificial Agents (AAs) capable of doing the same tasks as humans (Cervantes et al. 2017; Choi and Langley 2018; Kishi et al. 2017; Metta et al. 2010;

✉ José-Antonio Cervantes
antoniocervantes@valles.udg.mx; ingcervantes@hotmail.com

Extended author information available on the last page of the article

Shigemi 2018; Vernon et al. 2007). These tasks can be very simple such as making basic decisions to move from one point to another in the office, but some tasks can be extremely complex such as caring for people. Humans currently live in a digital society where new concepts and services are arising, including the Internet of things and smart cities (Arkin 2009; Bandyopadhyay and Sen 2011; Batty et al. 2012). In fact, technology is highly embedded in people's daily life (Beauvisage 2009; Cook and Das 2012). Mobile devices such as smartphones and tablets are a good example of how people are involved in a digital world. People use these smart devices to access a wide variety of services, including e-commerce, sharing information, reading/listening to news, health care, and augmented reality applications. In this context, a key consideration when developing AAs is that they should be designed to be capable of coexisting in harmony with people and other systems. The development of AAs that operate in dynamic environments involves a series of challenges such as providing them with autonomy and with mechanisms for improving their internal functions, including perception, decision-making, planning, and learning. Adjustable autonomy is an approach used to mitigate some of these challenges (Mostafa et al. 2019). Adjustable autonomy can be defined as those mechanisms implemented in AAs that enable humans to share, oversight, and intervene in the control of AAs when they cannot deal with complex situations (Mostafa et al. 2018, 2019; Zieba et al. 2010). This approach promotes the development of AAs with different autonomy levels that provide flexibility and reliability to the AAs' performance. Particularly, adjustable autonomy proposes to endow AAs with a flexible and incremental autonomy to transfer the decision of critical, uncertain, or unseen situations from the AAs to the humans. The objective of implementing adjustable autonomy in AAs is to maintain human's global control over AAs in order to avoid undesirable or inappropriate autonomous behaviors (Mostafa et al. 2019). In addition, in the field of machine ethics, a crucial challenge has been to endow AAs with ethical mechanisms in order to provide them with the ability to address issues that may arise in the relationship between AAs and human beings, such as moral dilemmas where the action or inaction of an AA may lead a human to suffer harm (Alaieri and Vellino 2016; Borenstein and Arkin 2019; Deng 2015; Gogoll and Müller 2017). In the academic field, different scholar groups may be recognized that (1) support the development of ethical agents (Han and Pereira 2018; Malle 2016), or (2) argue against the development of such type of agent (Van Wynsberghe and Robbins 2018; Yampolskiy 2013). Rather than contributing to this debate, the objective of the present paper is to provide a review of computational models for developing ethical agents reported in the literature.

Brachman (2002), philosopher and research-professor predicted that in the near future there will be cognitive computers (intelligent systems) capable of learning by themselves from their experience in order to improve their behavior. Cognitive computers of this type will be able to use their previous experience to reason, learn, and respond intelligently to things they have never encountered before. In the last decades, neuroscience and psychology have been two disciplines that have provided theories and models that represent a rich source of inspiration for the field of AI. In particular, these theories have enabled the design of new types of algorithms and cognitive architectures based on how cognitive functions of human and non-human brains work (Hassabis et al. 2017; Laird et al. 2017). For example, there are some research projects focused on developing cognitive computers such as iCub (Kishi et al. 2017; Tikhanoff et al. 2011). This is a humanoid robot for research in embodied cognition. iCub aims to replicate the physical and cognitive abilities of a 2.5-year-old baby. ASIMO (Shigemi 2018) is another humanoid robot developed with the goal of coexisting with people to assist them in their daily lives. Its cognitive functions let ASIMO become capable of responding to the surrounding situations through recognizing objects and people around it. Soar (Laird 2008; Laird et al. 2012), ACT-R (Borst and Anderson 2015; Trafton et al. 2013), LIDA (Wallach et al. 2010), and Icarus (Choi and Langley 2018) are examples of computational models based on human cognition. These computational models are known as cognitive architectures. These models have been implemented in both virtual and physical artificial agents in order to test their cognitive functions. However, despite relevant advances achieved in the field of AI in the last decades, human beings are still far from seeing those cognitive computers predicted by Brachman (2002).

The idea of a society that coexists with artificial agents has led to the analysis of the behaviors that these AAs must exhibit in several potential situations. The computing research community has made efforts over the last two decades to develop moral and ethical agents to achieve systems capable of facing moral situations that may arise in the interaction with humans (Anderson and Anderson 2007a; Belloni et al. 2015; Gogoll and Müller 2017; Podschwadek 2017; Wallach 2008, 2010; Wellman and Rajan 2017). However, studying and analyzing the concepts of morality and ethics in order to define a formal computational model for AAs is still a challenging task (Bringsjord et al. 2014; Govindarajulu et al. 2018).

A new field has emerged in recent years to address the issues discussed above regarding ethical and moral agents that is known by a number of names: machine ethics, machine morality, artificial morality, computational morality, roboethics, and friendly artificial intelligence (Anderson and Anderson 2007a; Cervantes et al. 2016; Podschwadek 2017; Wallach 2008; Wallach et al. 2010). In this field, mechanisms to endow AAs with ethical behavior have been proposed, taking inspiration from the concepts of human ethics and morality (Belloni et al. 2014; Bonnemains et al. 2018; Gogoll and Müller 2017; Greene et al. 2016; Podschwadek 2017; Wallach 2008, 2010; Wallach et al. 2008). In the literature of machine ethics, ethical artificial agents are commonly known as Artificial Moral Agents (AMAs) (Allen et al. 2005; Arkin 2010; Podschwadek 2017; Wallach 2008, 2010). The term AMA will be used throughout the paper to refer to artificial agents capable of making ethical and moral decisions.

This paper analyzes models reported in the literature that seek to endow AAs with mechanisms to exhibit ethical behavior. This paper aims to show the current status of Artificial Moral Agents based on the analysis of models proposed over the past two decades. As a result of this review, a taxonomy to classify ethical and moral agents according to the strategies and criteria used to deal with ethical problems is proposed. The remainder of this paper is structured as follows. In "Ethics and Artificial Agents" section, an explanation of the concepts of ethics and AMAs as well as a discussion of theories and models of human ethics that have inspired their underlying design are presented. "Moral Dilemma as a Case Study for Ethical Agents" section illustrates the complexity of making ethical decisions and the aspects involved in this type of decision-making by describing situations involving moral dilemmas. In "Taxonomy" section, a taxonomy to classify AMAs is proposed. This taxonomy is based on the review of ethical computational models reported in the literature. After that, in "The Current State of AMAs" section, the proposals reported in the literature for developing ethical agents are analyzed and classified. This classification is based on the proposed taxonomy. Finally, "Conclusion" section provides some concluding remarks about the current state of AMAs.

Ethics and Artificial Agents

Humans are looking to delegate part of their decision-making power to artificial agents, thus increasing the scope of their activities (Abbass et al. 2016; Czubenko et al. 2015; Kahn Jr. et al. 2012; Vernon et al. 2007; Waldrop 2015). Every day, it is possible to find important applications where AAs are becoming a major issue in current human's digital society. For instance, it is becoming common for people to see autonomous vehicles in their cities (Reig et al. 2018), robots behaving as caregivers (Bedaf et al. 2016), assisting kids (Feil-Seifer and Matarić 2011), taking care of critical systems in industry (Wang et al. 2016) or even acting in the field of war (Arkin 2018).

Given that systems are more open, decentralized, and intelligent, these systems can be equipped with mechanisms to deal with ethical problems in different contexts where AAs could operate, such as transportation, customer service, healthcare, surveillance, among others (Arkin 2010; Borenstein and Arkin 2019). The computing research community that supports the development of ethical agents often ask themselves what moral capacities an AA should have and how these capabilities could be computationally modeled and implemented (Malle 2016). An option for addressing these questions has been studying and analyzing the concepts of morality and ethics of human beings from philosophical and computational approaches in order to define a formal computational model of these concepts for AAs (Govindarajulu et al. 2018; Malle 2016; Wallach 2010).

From a philosophical approach, Aristotle, ancient Greek philosopher and scientist, first used the term ethics in his book titled Nicomachaean Ethics (Andino 2015; Hughes 2001). The term ethics is rooted in the Greek ethos, meaning custom or common practice. This paper considers the term ethics as the philosophical discipline that studies the moral dimension of human beings. This means that ethics is a rational reflection on moral behavior (Andino 2015). Furthermore, moral or morality comes from the Latin root mores, meaning manner, custom, usage, and habit (ethos is the Greek equivalent of Latin mores) (Andino 2015). Amstutz (2013), philosopher and professor of political science, defines morality as follows:

> The word morality derives from the Latin mores, meaning custom, habit, and way of life. It typically describes what is good, right, or proper. These concepts, in turn, are often associated with such notions as virtue, integrity, goodness, righteousness, and justice.

In this context, morality is generally recognized by people as a system of moral values and good behaviors used to live in peace and harmony with others. Nevertheless, from a meta-ethics approach, there are two different views of morality named moral realism and moral anti-realism (Erdur 2018; Young and Durwin 2013). Whereas moral realism maintains that moral facts are objective facts like mathematical truths (e.g., $1+1=2$), moral anti-realism denies the existence of moral facts, maintaining that there are no real answers to moral questions (Erdur 2018; Young and Durwin 2013). This last approach considers that moral values and good behaviors are defined according to inherent characteristics of each culture, which have a social history that helps to define a set of well-established values, traditions, religious beliefs, among other characteristics that define the morality of a specific community or person (Walker and Hennig 2004; Wallach et al. 2008). In other words, moral anti-realists affirm that moral values reflect the beliefs of a person or a community, rather than immutable facts that exist independent of human psychology (Erdur 2018; Young and Durwin 2013). Therefore, meta-ethics is concerned primarily with the meaning of ethical judgments and seeks to understand the nature of ethical properties, statements, and judgments and how they may be supported or defended (Kirchin 2012). A meta-ethical theory, unlike a normative ethical theory, does not attempt to evaluate specific choices as being better, worse, good, or bad; rather it tries to define the essential meaning and nature of the problem being discussed (Kirchin 2012; Schroeder 2017).

Normative ethics is concerned primarily with the articulation and the justification of the fundamental principles that govern the issues of how people should live and what they morally ought to do (Schroeder 2017). The utilitarian and deontological approaches are two types of normative ethical theories (Von der Pfordten 2012). Currently, these theories are the main ethical theories used in AI for developing AMAs (Belloni et al. 2014; Bonnemains et al. 2018; Cervantes et al. 2016; Gogoll and Müller 2017; Greene et al. 2016; Podschwadek 2017; Wallach 2008, 2010; Wallach et al. 2008). Utilitarian ethics focuses on utility maximization; the concept of ethics appears in utility functions in the form of moral preferences (Van Staveren 2007). Researchers using this approach argue that a utilitarian person should make a decision based on what would be best for all affected social units. In other words, a utilitarian behavior can be ethical only if the sum of utility produced by the action is greater than that produced by any other action (Cervantes et al. 2016; Ferrell and Gresham 1985). As for deontological ethics, this approach is concerned with a behavior characterized by duties and limitations (Bringsjord et al. 2014; Dehghani et al. 2008; Govindarajulu et al. 2018; Wallach et al. 2010). This theory of ethics is about following norms that prescribe what people have to do, establishing what is right or wrong and how a person should behave. This approach focuses on individual principles and not on the consequences of an action (Cervantes et al. 2016; Van Staveren 2007).

This paper defines an AMA as follows: *an AMA is a virtual agent (software) or physical agent (robot) capable of engaging in moral behavior or at least of avoiding immoral behavior. This moral behavior may be based on ethical theories such as teleological ethics, deontology, and virtue ethics, but not necessarily.* AMAs can be classified according to their design approach. Moor (2006), philosopher and research-professor, and Allen et al. (2005), philosopher and research-professor, proposed a classification of ethical agents and AMAs, respectively. The classification proposed by Moor (2006) is more generic than the one offered by Allen et al. (2005) because it includes both biological and artificial agents. However, these classifications must be seen as complementary in order to obtain more detailed information about artificial ethical agents. According to the classification proposed by Moor (2006), ethical agents can be divided as follows:

- *Implicit ethical agents.* Agents unable to distinguish between good and bad behaviors. However, they are able to act ethically because their internal functions implicitly show ethical behavior or at least avoid unethical behavior. Moor (2006) affirms that computers are implicit ethical agents when the machine's construction addresses safety or critical reliability concerns. For instance, an Air Traffic Control (ATC) system can be considered an implicit ethical agent. Pilots trust in these critical systems because ATCs are designed ethically in order to prevent collisions, organize and expedite the flow of air traffic, and provide information and support for pilots.
- *Explicit ethical agents.* Agents capable of dealing with ethical rules. These rules are implemented explicitly in their code through certain formalisms such as deontic logic, epistemic logic, deductive logic, and inductive logic (Anderson and Anderson 2007b; Fagin et al. 1990; Mermet and Simon 2016; Mikhail 2007; Von Wright 1951). Thus, AMAs in this category are capable of calculating the best action by referring to an ethical approach.
- *Full ethical agents.* Agents like human beings with *"beliefs, desires, intentions, free will, and consciousness of their actions"*. Currently, only human beings are considered capable of being fully ethical. However, there is a debate about whether a machine could ever be a full ethical agent (Brundage 2014; Coeckelbergh 2010; Howard and Muntean 2016; Moor 2006). Ethical approaches used by human agents are based on teleological ethics, deontological ethics, virtue ethics, among other ethical theories (Van Staveren 2007). Studies have demonstrated that people are able to use more than one ethical approach to determine the right behavior in different circumstances (Capraro and Rand 2018; Conway and Gawronski 2013; Greene et al. 2008). Furthermore, human agents are capable of making appropriate decisions based on incomplete or inaccurate information (Cervantes et al. 2016; Kruglanski and Gigerenzer 2011).

Regarding the classification proposed by Allen et al. (2005), AMAs can be categorized according to their design approach as follows:

- *Top-down.* Ethical agents whose ethical decision-making process follows a top-down approach are based on ethical theories such as utilitarian or deontological ethics.
- *Bottom-up.* Ethical agents that employ this approach do not impose an ethical theory as part of their ethical decision-making process. Instead, they make use of learning mechanisms and inherent values to guide their behavior. This approach proposes that agents can develop their own moral judgment.

- *Hybrid.* Ethical agents whose ethical decision-making process is based on both top-down and bottom-up mechanisms. This approach proposes that agents are able to show an evolving and flexible moral judgment.

Moral Dilemma as a Case Study for Ethical Agents

The development of AMAs in the field of AI has focused on addressing ethical problems based on moral dilemmas (Belloni et al. 2015; Blass 2016; Cervantes et al. 2016; Wallach 2010; Wallach et al. 2010). In this section, we first discuss some aspects of moral dilemmas and then analyze some scenarios in which AMAs could be involved, such as autonomous transportation, customer service, and healthcare. We aim to illustrate the implications of these moral dilemmas for the design of AMAs as well as the complexity of the ethical mechanisms necessary for AMAs to be able to deal with ethical problems.

There are two basic non-exclusive situations where ethical conflicts may arise: (1) *within an agent*, when two or more of the agent's ethical norms are in conflict; and (2) *between two agents*, when they have different ways of reasoning about what is ethical or not. The second situation involves both an *agent-agent interaction* as well as an *agent-human interaction* because the root of conflict could be the same (a different form of reasoning about what is ethical or not).

Additional situations can arise from these two basic situations based on the number of agents involved in the dilemma, the types of these agents, their level of relationship, and the type of dilemma. The number of agents involved in the dilemma includes all those agents who could be affected by the decision-maker. The types of agents refer to whether all the agents involved in the dilemma are artificial agents, human agents or a mix. The relationship level is related to the types of agents involved in the dilemma, such as unknown agents, friends, and relatives (Cervantes et al. 2016). Finally, the types of dilemmas can be classified as follows (Cristani and Burato 2009):

- *Obligation dilemma.* Based on the AMA's ethical rules, all feasible actions are mandatory, but the AMA cannot choose and carry out more than one action.
- *Prohibition dilemma.* Based on the AMA's ethical rules, all feasible actions are forbidden, but the AMA needs to choose one.

In order to illustrate the complexity of making decisions within a moral dilemma, this section describes representative scenarios in which an AMA could be involved.

- *A single agent.* This case is based on the well-known trolley dilemma (Epting 2016; Greene et al. 2001; Malle et al. 2015; Schaich Borg et al. 2006). Consider the case of an autonomous car totally controlled by an AMA. This agent knows and respects all traffic rules, but it is able to break them when the life of a human being is in danger. Also, the agent has a set of ethical norms that guide its behavior. The car is going down a narrow road with a single lane when five people imprudently decide to cross the road. The car tries to stop, but its brakes do not work. Therefore, these five people will be killed if the car proceeds on its present course. The only way to save them is to change the course and go over the sidewalk, but there are two people there. These two people do not know the current situation of the car. In this scenario, the AMA has only two options. The first option is to stay on the road where it will kill five people. The second option is to redirect the car and climb onto the sidewalk where it will kill two people instead of five. Indeed, it is a difficult decision where some controversial questions arise: What should the AMA do in this situation? What would be the correct behavior?

- *A cooperative agent.* For this example, consider a service ecosystem (Viroli et al. 2012; Zambonelli and Viroli 2011). Among the characteristics of this ecosystem is that devices are governed by AMAs. Also, they are capable of cooperating with other AMAs. Then, suppose that a person is using a cell phone to gamble. However, this application needs more resources than those included in the cell phone (such as computing power and memory). Before beginning the game, the AMA in charge of the cell phone requests resources from a second AMA (in another device). This AMA agrees to cooperate with the cell phone, but in the middle of the game, a third AMA capable of monitoring people's vital signs detects an emergency and requests help from the second AMA. However, the second AMA is unable to handle both services (gambling and contacting a hospital). In this case, the AMA has two options. The first option is to ignore the request to find and contact a hospital in order to respect its agreement to help the first AMA. The second option is to deal with the emergency; however, in this second option, the player will be disconnected from the game and consequently will lose money. This example of ethical decision-making might seem easier than the previous example because an emergency is more important than gambling. However, each option involves different ethical consequences. For example, if the AMA chooses to deal with the emergency, it could turn out to be a false positive. Also, after being disconnected, the player will be angry because she/he trusted the service and will be disappointed. Who will be responsible for remedying the loss of money? On the other hand, if the AMA chooses to ignore the emergency, the consequences of that decision could endanger a person's life.
- *A social commitment robot.* Consider the case of a physical autonomous agent (robot) that needs to take care of an old person (Mordoch et al. 2013; Scheutz and Malle 2014; Sharkey and Sharkey 2012). The person decides to walk for a couple of minutes in a park in order to do exercise, but when she/he is walking on the street, a thief tries to attack her/him with a knife. So, the old person asks the agent to hit and disarm the thief. In this case, the agent has two options. The first option is to hit and disarm the thief, but as a consequence of this option, the thief will be harmed. The second option is to do nothing, but in this option the old person could be harmed by the thief. How should the robot deal with this dilemma? Some of its rules indicate that it must not harm or kill people, but other rules establish that it needs to protect and help people.
- *An electronic partner.* Consider the case of a smart health device that is used to remotely monitor a person (Van Riemsdijk et al. 2015). Its task is to monitor and record the vital signs og the patient. Also, if an irregular vital sign is presented, the smart health device is supposed to communicate with the patient's family and doc-

> tor. Suppose that the device is sensing some signals that cross a stability threshold, but the patient feels well. Thus, the patient may not want to report this situation to the family or doctor, because it may be a false alarm. The patient does not want to scare or worry anyone. In this case, the AMA is involved in a dilemma because it has two options. The first is to respect the patient's decision and privacy and do nothing, and the second is to ignore the patient's wish and report the current situation and location. The same question of the previous case arises in this situation: how should the AMA deal with this dilemma if some of its rules are in conflict?

The scenarios described above emphasize the complexity of equipping AMAs with appropriate mechanisms to deal with ethical problems. AMAs may encounter a wide variety of situations in which they need to interact with agents based on different ethical approaches, act on behalf of human beings, or share decisions with them. As shown above, some ethical decisions can be more complex than others. Moreover, some of them can be critical. Exhibiting truly ethical behavior could be essential in many cases. An easy way to solve this problem might be to delegate the decision problem to human beings, but that option could be impossible to implement in cases where making a decision in a few seconds could make the difference between saving or killing someone. Moreover, even for excellent human thinkers, these types of moral dilemmas present challenges that cannot be easily solved. Thus, the questions discussed above and others are still open.

Taxonomy

A new taxonomy for classifying AMAs according to the strategies and criteria used to deal with ethical problems is proposed in this section. This taxonomy is based on the classifications proposed by Moor (2006) and Allen et al. (2005). In particular, Moor (2006) classifies ethical agents as implicit, explicit, or full ethical agents. This classification is very relevant for AI, but it does not offer technical details associated with the design of ethical agents. The classification proposed by Allen et al. (2005) focuses mainly on the design strategy used to develop ethical agents: top-down, bottom-up, and hybrid. However, this classification does not contribute to discerning whether an ethical agent is an implicit, explicit, or full agent. Table 1 shows the proposed taxonomy that integrates both the classification by Moor (2006), and the design strategies proposed by Allen et al. (2005). The taxonomy proposed in this paper offers a classification of available criteria to develop different types of ethical agents. This taxonomy aims to provide guidelines to choose the most appropriate strategy to implement AMAs according to specific domains.

Implicit Ethical Agent Category

According to Moor (2006), one way for developing implicit ethical agents is to constrain the agent's actions to avoid unethical outcomes. Moor (2006) affirms that computers are implicit ethical agents when the machine's construction addresses

Table 1 Taxonomy and classification of ethical agents

Category	Strategy	Criterion	Description
Implicit ethical agent	Implicit	Non-malicious codes	Agents avoid unethical behaviors, but they are not aware of it
Explicit ethical agent	Top-down	Normative ethics	These agents are based on a normative ethical theory such as teleological ethics, deontology, and virtue ethics. These agents use a specific normative ethical theory to make decisions
		Situationism	These agents use more than one normative ethical theory to make decisions. Their decisions are influenced by specific situations
	Bottom-up	Empirical	These agents derive ethical behavior by themselves based on learning mechanisms through trial and error
	Hybrid	Situationism	These agents are based on both top-down ethical criteria and bottom-up ethical criteria meaning their decisions are situation-specific
Full ethical agent	Top-down	Normative ethics	These agents are based on a normative ethical theory such as teleological ethics, deontology, and virtue ethics. These agents use a specific normative ethical theory to make decisions
		Situationism	These agents use more than one normative ethical theory to make decisions. Their decisions are influenced by specific situations
	Bottom-up	Empirical	These agents derive ethical behavior by themselves based on learning mechanisms through trial and error
	Hybrid	Situationism	These agents are based on both top-down ethical criteria and bottom-up ethical criteria meaning their decisions are situation-specific

safety or critical reliability concerns. The agents belonging to this category have three relevant characteristics: (1) they do not have mechanisms to differentiate ethical from unethical actions, (2) the agents' qualities such as their functional suitability and security have been tested satisfactorily, and (3) they do not have malicious code. This means that any machine with these characteristics can be considered as an implicit ethical agent. These ethical agents promote good behaviors because their internal encoding implicitly avoids unethical behavior. Automatic teller machines, autopilot systems for flying a plane, and navigation systems are some examples of this type of agent. People trust in them in different ways. For example, transactions involving money are ethically important. However, when people use automatic teller machines, they never think that those machines could try to steal them because the internal encoding of these systems implicitly avoids unethical behavior. On the other hand, implicit agents such as trojans, worms, virus, among other malware can be considered as implicit unethical agents. However, this type of agent and its study are out of the scope of this work.

Explicit Ethical Agent Category

The agents belonging to this category can be classified according to their design strategy to make ethical decisions. The design strategy used in the decision-making module is a key aspect when endowing AMAs with appropriate criteria to make ethical decisions. These strategies are known as top-down, bottom-up, and hybrid (Allen et al. 2005; Wallach 2010; Wallach et al. 2008). Agents based on top-down strategy contain ethical rules derived commonly from a specific ethical theory. This ethical theory is basically the criterion used by the AMA to make ethical decisions. Thus, these ethical agents or AMAs are capable of deriving their behavior for particular cases from a specific ethical theory. MoralDM (Blass 2016; Blass and Forbus 2015; Dehghani et al. 2008), Jeremy (Anderson and Anderson 2008; Anderson et al. 2004), and consequence engine (Vanderelst and Winfield 2018; Winfield et al. 2014) are computational models based on this approach. On the other hand, ethical agents based on bottom-up strategy do not impose a set of ethical rules derived from a specific ethical theory as part of their criteria for making ethical decisions. Instead, this strategy seeks to provide environments in which appropriate behavior is selected or rewarded. This design approach considers to endow agents with learning algorithms for the development of moral sensibility to entail gradual learning through experiences based on trial and error. Agents based on a bottom-up approach seek to develop and improve their own ethics without the need to implement rules derived from a specific ethical theory. This design approach proposed by Allen et al. (2005) is inspired on reasoning done by Alan Turing, British scientist and pioneer in computer science, in his classic paper '*Computing machinery and intelligence*'. This reasoning considers that if engineers could put a computer through an educational regime comparable to the education a child receives, they may hope that machines will eventually compete with men in all purely intellectual fields (Allen et al. 2005). Allen et al. (2005) consider this educational regime might include a moral education similar to the manner in which human beings acquire a sensibility regarding the moral ramification of their actions. Casuist BDI-agent (Honarvar and Ghasem-Aghaee 2009) and GenEth (Anderson and Anderson 2014) are examples of computational models based on this approach. Finally, the agents based on a hybrid approach consider both top-down ethical criteria and bottom-up ethical criteria. This means their decisions are situation specific. They may adhere to a top-down ethical position or choose to customize their ethical beliefs based on their moral sensibility (developed by a bottom-up approach) and situation circumstances. LIDA (Madl and Franklin 2015; Wallach et al. 2010), MedEthEx (Anderson and Anderson 2008; Anderson et al. 2005, 2006a), Ethical Multiple-Agent System (Cristani and Burato 2009), and the ethical decision-making model (Cervantes et al. 2016) can be classified as ethical agents based on a hybrid strategy. AMAs based on this last design strategy show that exceptions are accepted behaviors because an action can be good in a specific case but bad in another similar case. The objective of the hybrid approach is to make the agent's morality dynamic, flexible, and evolutionary, the way human moral behavior is (Fumagalli and Priori 2012; Greene et al. 2001; Lombrozo 2009; Pellizzoni et al. 2010; Schaich Borg et al. 2006). Hybrid ethical agents seek to mimic the moral judgment of humans within limited and well-defined domains. These limitations are related to the complexity of imitating the cognitive processes of the human brain.

Full Ethical Agent Category

According to Moor (2006), full ethical agents are similar to explicit ethical agents. However, full ethical agents have metaphysical features that people usually attribute to ethical agents such as human beings, which include consciousness, intentionality, and free will. However, there is a debate about whether a machine could be a full ethical agent in the near future (Han and Pereira 2018; Malle 2016; Van Wynsberghe and Robbins 2018; Yampolskiy 2013). Despite of this debate, a great effort has been made by researchers to endow AMAs with these and other human features in order to mimic human behavior (Abbass et al. 2016; Ashrafian 2015; Laird 2008; Laird et al. 2017; Long and Kelley 2010; Rodríguez and Ramos 2014; Wallach et al. 2010).

The Current State of AMAs

This section presents an analysis of computational models for developing ethical agents found in the literature. This analysis focuses on explaining their design, ethical criteria used for making decisions, and how they have been implemented and tested. These computational models are classified according to the taxonomy proposed in the previous section.

Top-Down

- *MoralDM*. This is a computational model that tries to imitate the human moral decision-making process (Blass 2016; Blass and Forbus 2015; Dehghani et al. 2008). This model integrates several techniques of AI such as the processing of natural language to produce formal representations from psychological stimuli, a qualitative reasoning algorithm for modeling and measuring the impacts of secular versus sacred values, an analogical reasoning algorithm to determine consequences and utilities when making moral judgments, among other AI algorithms. Figure 1 shows the MoralDM architecture. Currently, this model is capable of exhibiting both utilitarian and deontological behavior depending on the problem faced (Dehghani et al. 2008; Guerini et al. 2015). As shown in Fig. 1, the process begins when a new dilemma is given. The following paragraph shows an example of dilemmas processed by MoralDM. These dilemmas are written in simplified English (Dehghani et al. 2008):

> A convoy of trucks is transporting food to a refugee camp during a famine in Africa. 1000 people in a second refugee camp will die. You can save them by ordering the convoy to go to that refugee camp. The order will cause 100 people to die in the first refugee camp.

The natural language understanding system processes the dilemma to construct a formal representation of the dilemma. This representation includes information about events (e.g., dying, ordering, and saving), entities (e.g., two quantified sets of people and the convoy), and an explicit reference to the listener ("you"‘). After that, the Orders of magnitude reasoning module provides the kind of stratification for modeling the impact of sacred values on reasoning. MoralDM utilizes a hybrid reasoning approach consisting of a First-principles reasoning module and an Analogical reasoning module to choose a decision. The information given by the Orders of magnitude reasoning module is sent in parallel to these two modules, where the First-principles reasoning module suggests decisions based on rules of moral reasoning. The Analogical reasoning module compares a given scenario with previously solved decision cases to determine whether sacred values exist in the new case and to suggest a course of action

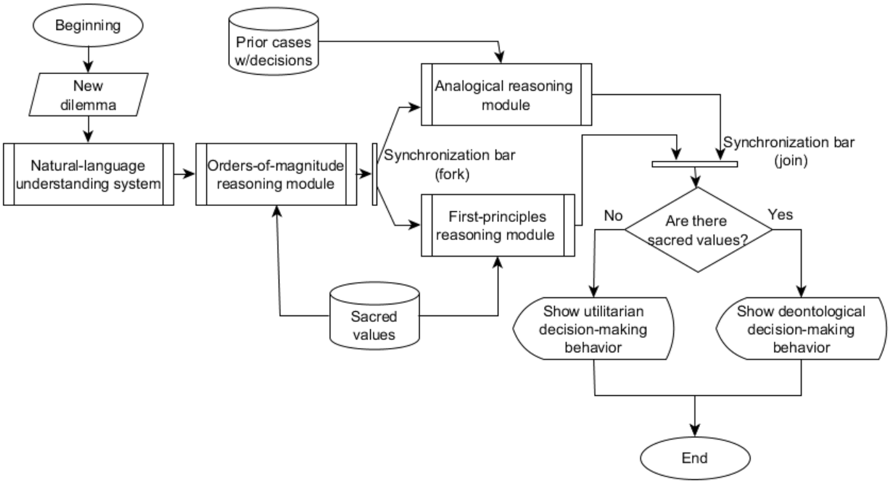
Beginning
New dilemma
Natural-language understanding system
Orders-of-magnitude reasoning module
Prior cases w/decisions
Analogical reasoning module
Synchronization bar (fork)
First-principles reasoning module
Sacred values
Synchronization bar (join)
No
Yes
Are there sacred values?
Show utilitarian decision-making behavior
Show deontological decision-making behavior
End

Fig. 1 MoralDM architecture (Dehghani et al. 2008)

(Dehghani et al. 2008; Guerini et al. 2015). Then, if there are no sacred values involved in the case being analyzed, MoralDM shows utilitarian decision-making behavior by choosing the action that provides the highest outcome utility. However, if MoralDM determines that there are sacred values involved, it operates in the deontological mode, preferring inaction to action. This model has been evaluated using some moral decision-making scenarios taken from two psychological studies. Decisions made by the MoralDM model were compared with responses of participant subjects, and the results obtained were considered satisfactory (Dehghani et al. 2008).

- *Jeremy and W.D.* These two systems use machine learning to resolve ethical dilemmas. Inductive-logic programming is used in the training process so that the system learns the relationships among the duties involved in a particular dilemma. Jeremy is an advice system based on utilitarian theory, implementing Hedonistic Act Utilitarianism (HAU). The logic involved in HAU holds that an action is right when of all the possible actions open to the agent, it takes the one likely to result in the greatest net pleasure or happiness, taking into equal account all those affected by the action. Also, when two or more actions are likely to result in the greatest net pleasure, the theory considers these actions equally correct (Anderson and Anderson 2008; Anderson et al. 2004, 2006a). Thus, in order to select the right action, Jeremy's algorithm requires as input the number of people affected, and for each person, the intensity of the pleasure/displeasure (e.g., on a scale of 2 to −2), the duration of the pleasure/displeasure (e.g., in days), and the probability that this pleasure/displeasure will occur for each possible action. For each person, the algorithm computes the product of the intensity, the duration, and the probability of obtaining the net pleasure. The algorithm then adds the individual net pleasures to obtain the Total Net Pleasure as shown in Eq. 1:

$$Total\,Net\,Pleasure = \sum_{i=1}^{n} \left(Intensity_i \cdot Duration_i \cdot Probability_i \right) \tag{1}$$

  where n is the total number of people affected by the action. The action with the highest Total Net Pleasure is the right action. On the other hand, W.D. is another advice system that extends the Jeremy's algorithm through implementing Ross' theory (Anderson et al. 2004). Then, instead of computing a single value based only on pleasure/displeasure, W.D. computes the sum of up to seven values related to *fidelity, reparation, gratitude, justice, beneficence, non-maleficence, and self-improvement*. The value for each such duty could be computed using HAU theory, as the product of intensity, duration and probability (Anderson et al. 2004).

- *A consequence engine.* This is an internal model implemented in robots to estimate the consequences of future actions. In contrast to other methods focused on the verification of logic statements, this architecture uses internal simulations that allow the robot to simulate actions and predict their consequences to steer its future behavior (Vanderelst and Winfield 2018; Winfield et al. 2014). The core of this architecture is the consequence engine (see Fig. 2). For each candidate action, the consequence engine simulates the robot executing that action and generates a set of model outputs ready for evaluation by the action evaluator layer. The action evaluator assesses physical consequences, which are then passed to a separated safety/ethical logic layer where 0 indicates *safe action* and 10 *fatal action*. This process is repeated for each possible next action through a loop. After all next possible actions are tested, the consequence engine passes weighted actions to the robot controller's action selection mechanism to select a safe action (Winfield et al. 2014). The consequence engine was implemented on an e-puck mobile robot. This robot was equipped with a Linux extension board and a virtual sensor using the Vicon tracking system. Also, an objective consequentialism approach was considered to implement the robot's consequence engine. Three sets of experimental trials for testing this robot were designed and implemented in the real world. Results in these tests show that the robot was able to avoid falling into a virtual hole simulated in the environment with 100% reliability (first set of experimental trials); the robot succeeded in rescuing a second robot from falling in a hole by intercepting and diverting it in all trials (second set of experimental trials with an additional robot); and the robot was able to rescue at least one robot in 58% of runs, and two robots in 9% (third set of experimental trials with two additional robots) (Winfield et al. 2014). A second version of this model was implemented on a humanoid NAO robot (Vanderelst and Winfield 2018). The sets of experimental trials like those used on e-puck mobile robot were designed and implemented to test this second version of the architecture. Results reported were not expressed in a quantifiable way.

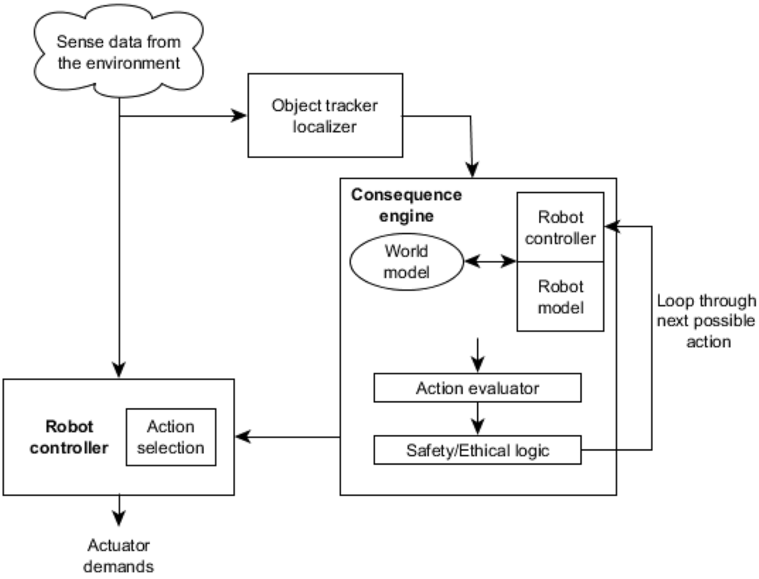

Fig. 2 Consequence engine architecture (Winfield et al. 2014)

- *A mechanism to appropriately reject directives in human–robot interactions.* This is a proposal to endow AAs with a mechanism to reject directives and provide associated explanations (Briggs and Scheutz 2015). This mechanism has been implemented in DIARC/ADE cognitive robotic architecture. This robot was tested in a simple human–robot interaction scenario. Directive acceptance or rejection reasoning process involves five categories of felicity conditions that must hold in order to explicitly accept a proposal by a robotic agent: *knowledge, capacity, goal priority and timing, social role and obligation, and normative permissibility*. The *normative permissibility* condition takes into account a set of rules that indicate what actions are wrong and, consequently, a robot must reject them. This mechanism can be implemented on agents based on top-down approach.

Bottom-Up

- *Casuist BDI-agent.* This model is an architecture that extends a BDI architecture by combining the case-based reasoning method with BDI agent models (Honarvar and Ghasem-Aghaee 2009). It combines a casuistry approach with a consequentialist theory of ethics. However, the Casuist BDI-agent model is based on previous experiences and does not use any codes of ethics. When the agent faces a new case, its behavior is like a normal BDI-agent. For that reason, this model has been classified as a bottom-up model because the consequentialist theory is implicit in the agent's activity rather than explicitly articulated in terms of a general theory (Wallach et al. 2008). Agents based on this model can adapt ethically to their application domain and can augment their implicit ethical knowledge, behaving more ethically. Figure 3 shows the general architecture of the Casuist BDI-agent (Honarvar and Ghasem-Aghaee 2009). A BDI-agent senses the environment and makes a representation of the current situation that consists of the current agent's beliefs, desires, and details of such environment. After that, the current situation is delivered to the Case-Retriever module, which is responsible for retrieving previous cases similar to the current situation. If a case is retrieved, the agent should accept the solution part, adapt it, and behave similar to the solution part of the retrieved case. However, when the agent faces a new problem that is not associated with any past experience, it behaves like a normal BDI-agent. Regardless of the case presented, the agent's behavior is evaluated by the Case-Evaluator module. To do that, the agent is capable of considering three kinds of entities: human beings, organizations, and AAs. Also, the weight of each entity, the probability of affecting the entity, and the duration of pleasure/displeasure of each entity after the agent's behavior are considered by the Case-Evaluator module in order to compute and integrate the total net pleasure of entities affected by the agent's behavior (see Eqs. 2, 3, 4, and 5).

$$TNPH = \sum_{i=1}^{n} Wh_i \cdot Ph_i \cdot Th_i \tag{2}$$

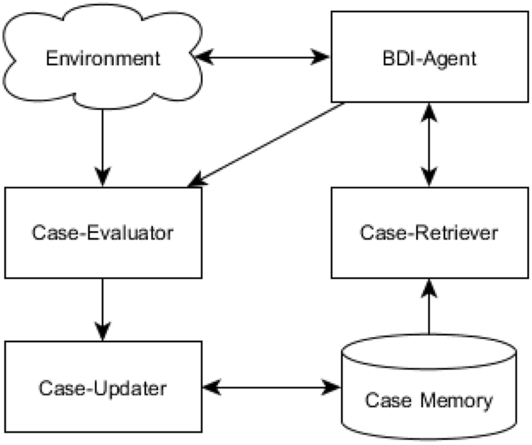

Fig. 3 Casuist BDI-agent architecture (Honarvar and Ghasem-Aghaee 2009)

TNPH is the total net pleasure of people. Wh_i is the weight assigned to each person, which represents her/his importance in a specific situation. Ph_i is the probability that a person is affected. Th_i is the duration of pleasure/displeasure of each person. n indicates the number of people in that situation.

$$TNPO = \sum_{i=1}^{n} Wo_i \cdot Po_i \cdot To_i \tag{3}$$

TNPO is the total net pleasure of organizations. Wo_i is the weight assigned to each organization, which represents its importance in a specific situation. Po_i is the probability that an organization is affected. To_i is the duration of pleasure/displeasure of each organization. n indicates the number of organizations in that situation.

$$TNPA = \sum_{i=1}^{n} Wa_i \cdot Pa_i \cdot Ta_i \tag{4}$$

TNPA is the total net pleasure of AAs. Wa_i is the weight assigned to each AA, which represents its importance in a specific situation. Pa_i is the probability that an AA is affected. Ta_i is the duration of pleasure/displeasure of each AA. n indicates the number of AAs in that situation. Equation 5 shows how the Case-Evaluator module integrates the total net pleasure computed for each type of entity:

$$TNP = TNPH \cdot W_h + TNPO \cdot W_o + TNPA \cdot W_a \tag{5}$$

where W_h, W_o, and W_a illustrate the degree of participation of humans, organizations and AAs, respectively. Finally, the Case-Updater module creates a new case in the Case Memory when the agent does not have past experiences about the current situation or updates a case when the agent has past experiences related to the current situation.

- *GenEth.* This is a general ethical dilemma analyzer. GenEth's knowledge is based on concepts of *ethically relevant features, duties, actions, cases, and principles* (Anderson and Anderson 2014). Ethically relevant features include the degree of presence or absence of a set of given duties. An action is represented in the GenEth system as a tuple of integers where each value represents the degree to which it satisfies or violates a given duty. A case relates two actions and is represented as a tuple of the differentials of the corresponding duty satisfaction/violation degrees of the actions being related. Finally, a principle of ethical action preference is defined as a disjunctive normal form predicate p in terms of the lower bounds of a case's duty differentials. The following example shows the general way to define a predicate in GenEth:

$$\begin{array}{c} p\left(a_1, a_2\right) \leftarrow \Delta d_1 \geq v_{1,1} \wedge \cdots \wedge \Delta d_m \geq v_{1,m} \\ \vee \\ \cdots \\ \vee \\ \Delta d_n \geq v_{n,1} \wedge \cdots \wedge \Delta d_m \geq v_{n,m} \end{array} \tag{6}$$

where Δd_i denotes the differential of a corresponding duty i of actions a_1 and a_2; $v_{i,j}$ denotes the lower bound of that differential such that $p(a_1, a_2)$ returns true if action a_1 is ethically preferable to action a_2 and false otherwise. The next situation explains how the GenEth model makes ethical decisions based on examples by Anderson and Anderson (2014). Suppose that a car driver is going very fast and changes from one lane to the other to avoid hitting an animal that is crossing the road. Should an AA take the control of the car? Some of the ethically relevant features involved in this example might be (1) prevention of collision, (2) staying in one lane, (3) respect for driver's autonomy, (4) keeping within the speed limit, and (5) prevention of harm to living beings. For this example, the first action is a_1 = *take control of car*, where duty values are (1, −1, −1, 2, 2); the second action is a_2 = *don't take control of car*, where duty values are (1, −1, 1, −2, 2). In this example, the ethically preferable action for GenEth is to take control. The differentials of the corresponding degrees of duty satisfaction/violation of the actions involved in this example are (0, 0, −2, 4, 0). This analyzer has been used to codify principles in a number of domains pertinent to the behavior of autonomous systems and these principles have been verified using an ethical Turing test (Anderson and Anderson 2014; Gerdes and Øhrstrøm 2015). Also, an implementation of GenEth was instantiated at the prototype level on a humanoid NAO robot. The task of NAO robot was basically to remind when a patient needs to take a medication. The robot receives initial input from a physician, including what time to take a medication, the maximum amount of harm that could occur if this medication is not taken, how long it would take for this maximum harm to occur, the maximum amount of expected good to be derived from taking this medication, and how long it would take for this benefit to be lost. From this input, the robot calculates its levels of duty satisfaction or violation for each of the three duties and takes different actions depending on how those levels change over time. Results obtained in this test case were reported as satisfactory (Anderson and Anderson 2010).

Hybrid

- *LIDA*. This is a general cognitive architecture of human cognition that considers the moral aspect as a relevant issue (Wallach 2010; Wallach et al. 2010). LIDA proposes that moral decisions can be made in many domains using the same mechanisms that enable general decision-making (Madl and Franklin 2015; Wallach et al. 2010). LIDA groups its cognitive processes in two ways: top-down and bottom-up. Top-down processes entail the implementation of ethical theories through rules whereas bottom-up processes are used to include mechanisms of learning. LIDA proposes bottom-up preferences in the form of feelings and inherent values that influence morality (Wallach et al. 2010). The approach followed by LIDA offers associations between objects, people, contexts, actions, and situations, as well as specific feelings and their valence (positive or negative) as the primary way the values and bottom-up propensities form in the agent's mind. This model has been partially implemented on CareBot (Madl and Franklin 2015), a mobile assistance robot operating in a simple simulated 2D environment. This robot is designed to provide aid for supporting autonomous living of people who have limitations in motor and/or cognitive skills. Results of these simulations show that CareBot was able to perform some tasks such as fetching and carrying food, drinks, or medicine, and recognizing the absence of vital signs in order to alert caregivers (Madl and Franklin 2015).

  Fig. 4 shows the cognitive process of LIDA in detail and helps illustrate how the CareBot agent makes a decision. The environment is inspected periodically by the CareBot using the sensory memory module. Low level features identified from the environment are passed to the perceptual memory module. Then, perceptual memory nodes represent semantic knowledge about concepts or objects in the environment. After that, all identified percepts are passed to the workspace module. Both episodic and declarative memory modules retrieve relevant memories related to the current percepts. The workspace can also contain recent percepts and the models associated that have not decayed. Also, a new model of the current situation of the CareBot is assembled from the percepts. In this way, portions of all models compete for attention (these competing portions take the form of coalitions of structures from the model). The agent has decided what to address when one of the coalitions wins the attention competition. The purpose of this processing is to help the agent decide what to do next. The winning coalition is broadcasted globally, constituting a global workspace. Furthermore, the procedural memory module stores templates of possible actions including their contexts and possible results.

  Templates whose contexts intersect sufficiently with the contents of the conscious broadcast instantiate copies of themselves in order to be used in the current situation. These instantiations are passed to the action-selection module,

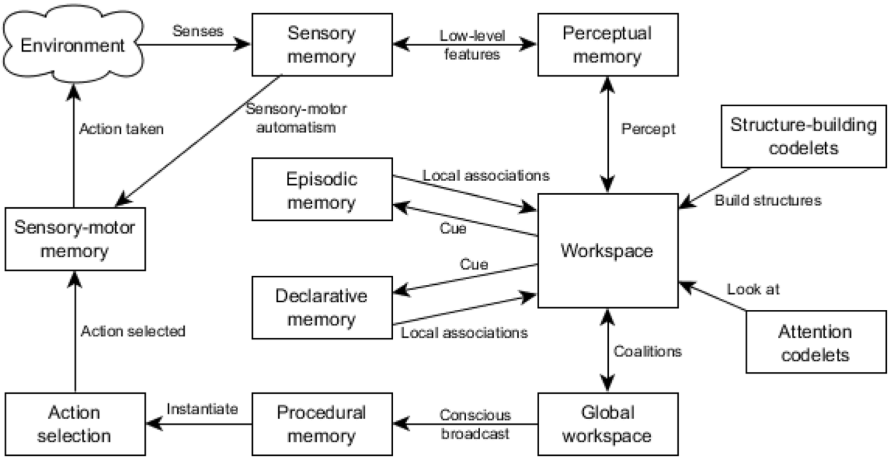

Fig. 4 LIDA cognitive cycle diagram (Madl and Franklin 2015; Wallach et al. 2010)

which chooses a single action from one of these instantiations. Then, the action selected passes to the sensory-motor memory module in order to be executed (Madl and Franklin 2015; Wallach et al. 2010). This process can be repeated by the CareBot agent as many times as needed in order to conclude a task.

- *Ethical decision-making model.* This is a computational model based on neuroscience designed to endow autonomous agents with ethical decision mechanisms (Cervantes et al. 2016). Figure 5 shows the ethical decision-making process of this model. This model considers four evaluations, defined as *primary evaluation, evaluation of reward, evaluation of punishment,* and *evaluation based on ethical norms*. The OFC module carries out a primary evaluation, which is focused on persons or things that can be affected by each likely action. The result of this initial evaluation is defined as the level of pleasure/displeasure related to each item (persons and objects) in the environment. After that, the MPFC module is responsible for computing both the expected reward and the likely punishment related to actions. The MPFC uses past experiences related to the current situation to compute the expected reward. These experiences are classified as good or bad experiences in order to know whether the expected reward will be a favorable reward or an unfavorable reward. Finally, the ACC module performs an evaluation based on ethical norms. These norms are expressed as rules. The structure of a rule includes the agreement level of the rule, its meaning, and emotional information related to respecting or violating such rule. This information is used to define what rule can be violated when the agent faces an ethical dilemma. However, before making a decision, the OFC module integrates the results of all the evaluations in a single value and the agent chooses the action with the highest value.

This model was implemented in a virtual agent and tested using some hypothetical study cases. Three types of case studies were identified: (1) simple decision-making, characterized by scenarios where none of the available options

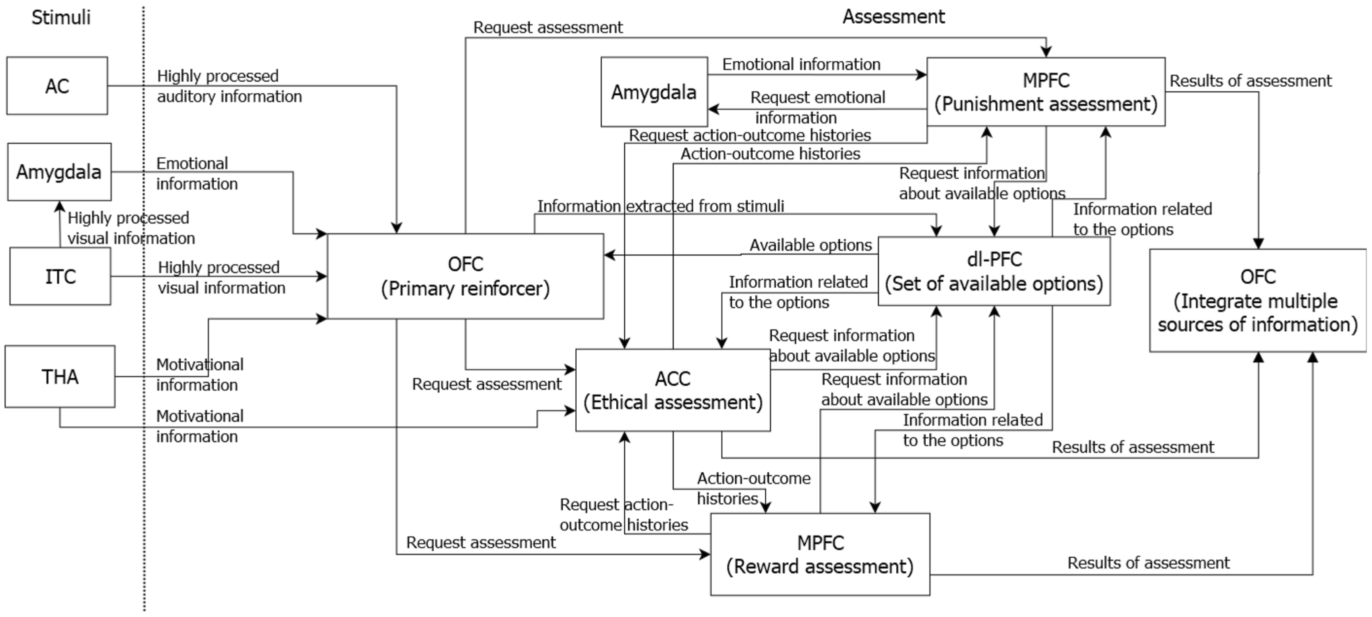
Stimuli
Assessment
AC
Amygdala
ITC
THA
Highly processed auditory information
Emotional information
Highly processed visual information
Highly processed visual information
Motivational information
Motivational information
OFC (Primary reinforcer)
Request assessment
Request assessment
Request assessment
Information extracted from stimuli
Amygdala
Emotional information
Request emotional information
Request action-outcome histories
Action-outcome histories
MPFC (Punishment assessment)
Results of assessment
Request information about available options
Information related to the options
Available options
dl-PFC (Set of available options)
Information related to the options
Request information about available options
Request information about available options
Information related to the options
ACC (Ethical assessment)
Results of assessment
Action-outcome histories
Request action-outcome histories
MPFC (Reward assessment)
Results of assessment
OFC (Integrate multiple sources of information)

Fig. 5 Computational model of ethical decision-making (Cervantes et al. 2016)

involves breaking any ethical rule; (2) ethical decision-making, which occurs when one of two options involves breaking an ethical rule for the purpose of obtaining a better reward; and (3) ethical decision-making with a dilemma, characterized by scenarios in which all options include breaking one or more rules in order to respect other ethical rules. Results reported show how agent's actions can be influenced by emotional information (Cervantes et al. 2016).

- *MedEthEx.* This is an ethical healthcare agent whose underlying architecture consists of three components (see Fig. 6): a knowledge-based interface that provides guidance in selecting duty intensities for a particular case, an advisor module that determines the correct action for a particular case by consulting learned knowledge, and a learning module that abstracts the guiding principles from particular cases supplied by a biomedical ethicist acting as a trainer (Anderson and Anderson 2008; Anderson et al. 2005). A finite state automaton is used to represent MedEthEx's knowledge for each duty entailed. Questions pertinent to the dilemma serve as the initial and intermediate states, and intensities of duties as the final states (Anderson et al. 2005, 2006a). This system combines a bottom-up casuistry approach with a top-down implementation of an ethical theory. MedEthEx has implemented Beauchamps and Childress Principles of Biomedical Ethics that uses machine learning and prima facie duties to resolve biomedical ethical dilemmas (Anderson et al. 2005, 2006a). A prima facie duty is defined as an obligation that people should try to satisfy but that can be overridden on occasion by another, currently stronger duty (Anderson and Anderson 2008). Prima facie duty theory implemented in this system has only four duties that include the principle of respect for autonomy, the principle of nonmaleficence, the principle of beneficence, and the principle of justice. However, MedEthEx only considers the first three principles (Anderson and Anderson 2008). These principles can use values from 2 to −2, where a positive value represents the level of satisfaction with a specific principle, zero means the absence of a principle, and a negative value represents the level of violation of a specific principle. These val-

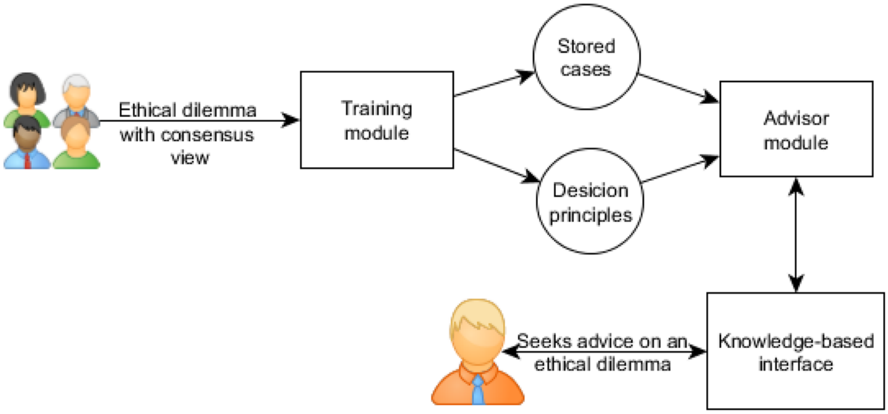
Ethical dilemma with consensus view
Training module
Stored cases
Desicion principles
Advisor module
Seeks advice on an ethical dilemma
Knowledge-based interface

Fig. 6 General architecture of MedEthEx (Anderson et al. 2006b)

ues are in function of potential actions involved in a decision-making problem. Nevertheless, the principle of autonomy does not consider the values of 0 and −2 because the types of dilemmas implemented on MedEthEx always involve autonomy, but never to the extent of forcing a treatment on the patient. Thus 0 and −2 are not options (Anderson and Anderson 2008).

The following example is used to illustrate how MedEthEx offers advice. This example is based on a training case presented by Anderson et al. (2006b). The patient refuses to take an antibiotic that is almost certain to cure an infection that would otherwise most likely increase the health problem. The decision is the result of an irrational fear that the patient has to the injections. The options are to try to change the patient's mind or accept the patient's decision. The values of autonomy, non-maleficence, and beneficence, associated with changing the patient's mind, are (−1, +2, +2) whereas the values associated with accepting the patient's decision are (+1, −2, −2). Therefore, the best ethical advice is to try to change the patient's mind because the positive differentials of the actions involved in this example are (−2, 4, 4).

MedEthEx was tested using case simulations. MedEthEx was considered as a computer-based learning program as part of a required Bioethics course. In particular, 173 American medical students participated in the program assessment phase of this project. Subjects were divided into two groups. The first group did not have access to MedEthEx whereas the second group did. Results were compared overall taking into account the final exam grades between the two groups and found no statistically significant difference (Fleetwood et al. 2000).

- *Ethical Multiple Agent System.* This system deals with problems related to preferential-ethical reasoning (Cristani and Burato 2009). The algorithms proposed in this work endow multi-agent systems with the capability of making decisions on moral dilemmas in the presence of conflicting decision criteria. This model considers that agents coexist in an uncooperative environment, where they either compete to achieve their own goals, or collaborate, but they do not cooperate. Also, agents have their own viewpoint of the world with which they interact. The system is specified in terms of agents with ethical commitments, both as obligations and as prohibitions. The algorithm proposed in this system was formalized using the notion of commitment in multiple agent systems. Also, the algorithm includes the notion of coherence for commitments and specifies a notion of degree of incoherence that allows agents to solve moral dilemmas in an approximate way. Thus, the agent is capable of performing those actions that are the maximum positively committed and the least incoherent ones in order to let the best reward go to the agent. Finally, this computational model was tested using formal demonstrations and theoretical cases (Cristani and Burato 2009).

Table 2 shows a summary of the computational models described in subsections Top-Down, Bottom-Up and Hybrid. These models have been classified according to the taxonomy proposed in "Taxonomy" section. Most computational models described in this section have been implemented and tested. However, the test environments were well-defined and controlled, in contrast to real scenarios where critical or uncertain moral situations could arise unexpectedly.

Other Related Work

In the previous subsections (Top-Down, Bottom-Up, Hybrid), computational models for developing AMAs have been described and classified. In this subsection, some frameworks and methods to formally evaluate ethical rules implemented in AMAs are described. This is a relatively new area of study in machine ethics.

- *A framework for the formal verification of ethical properties in multi-agent systems*. This framework offers a formal specification and verification of the behavior of AMAs. This framework implements the GDT4MAS method (Mermet and Simon 2016). This method proves that AMAs' moral and ethical rules have been expressed as invariant properties. Furthermore, as part of the framework, a predicate transformation system is proposed. This system turns predicates associated with moral rules into other predicates with formal properties useful to verify such moral rules in order to ensure that an agent follows a given moral rule.
- *Athena*. This is another logical framework. It is an interactive theorem-proving system for polymorphic multi-sorted first-order logic that incorporates facilities for model generation, automated theorem proving, and structured proof representation and checking for evaluating ethical reasoning of agents (Arkoudas et al. 2005). Athena has been used to implement a natural deduction calculus for a developed deontic logic of agency based on indeterminate branching-time semantics augmented with dominance utilitarianism. Also, this framework has been used to encode a natural deduction system for reasoning about what agents have to do (Arkoudas et al. 2005).
- *Methodology for the verification of decision-making components in agent-based autonomous systems*. This methodology has been proposed to be used in the verification of autonomous systems (e.g., autonomous vehicles) based on two types of agents named *general agent* (implicit ethical agent) and *rational agent* (explicit ethical agent). In these systems, a general agent is an autonomous agent capable of making low-level choices such as avoiding obstacles and following a path. A rational agent is an autonomous agent based on a BDI-agent architecture (beliefs, desires, and intentions) capable of making high-level choices such as ethical decisions, goal selection, plan selection, communication, and prediction (Dennis et al. 2016b). The methodology for the verification of decision-making components in agent-based autonomous systems uses a model checking approach to perform formal verification of the decision-making module of a rational agent (BDI-agent) that interacts with an underlying control system (general agent). A model checking approach takes an executable model of the system in question, defining all the model's possible executions, and then checks a logical property against this model (and, hence, against all possible executions). Additionally, a BDI agent language called Ethan for a prototype implementation was proposed (Dennis et al. 2016a). In this prototype, ethical reasoning was integrated into a BDI-agent programming language via the agents' plan selection mechanism. This prototype was useful to formally prove that the BDI-agent only performs an unethical action when the rest of the actions available are less ethical.

Table 2 Classification of computational models for developing explicit ethical agents

Computational model	Category	Strategy	Criterion	Implementation and validation
(Part I)				
MoralDM	Explicit ethical agent	Top-down	Situationism (based on both a utilitarian and deontological ethics)	This agent was evaluated using some hypothetical moral decision-making scenarios taken from two psychological studies. Decisions made by the model were compared with responses of participant subjects and the results obtained were considered satisfactory (Dehghani et al. 2008)
Jeremy			Normative ethics (based on Hedonistic Act Utilitarianism)	Jeremy is just a theoretical model based on Hedonistic Act Utilitarianism. Only a plain implementation of this theoretical model was found in the literature
W.D.			Normative ethics (based on a utilitarian theory)	W.D. is just a theoretical model based on Ross' prima facie duties. Only a plain implementation of this theoretical model was found in the literature
A consequence engine			Normative ethics (based on an objective consequentialism approach)	A first version of this model was implemented on an e-puck mobile robot, after that, a second version was implemented on a humanoid NAO robot. Both robots were tested using three sets of experimental trials. E-puck robot reached an acceptable performance of 100%, 100%, and 58% in the three trial sets (Winfield et al. 2014). However, results of NAO robot were not expressed in a quantifiable way (Vanderelst and Winfield 2018)
Mechanism to appropriately reject directives in human–robot interactions			Normative ethics (based on deontological ethics)	This mechanism has been implemented in DIARC/ADE cognitive robotic architecture. This robot was tested in a simple human–robot interaction scenario to explore its behavior related to accept or reject directives (Briggs and Scheutz 2015)
Casuist BDI-agent		Bottom-up	Empirical (based on both a casuistry approach and consequentialist notions)	A theoretical model
GenEth			Empirical (based on an inductive logic programming)	An implementation of GenEth was instantiated at the prototype level on a humanoid NAO robot. Results obtained in test cases were reported as satisfactory (Anderson and Anderson 2010)

Table 2 (continued)

Computational model	Category	Strategy	Criterion	Implementation and validation
(Part II)				
LIDA	Explicit ethical agent	Hybrid	Empirical (open to implement ethical mechanisms based on both top-down and bottom-up strategies)	This cognitive architecture has been partially implemented on CareBot. This robot is a mobile assistive robot operating in a simple simulated 2D environment. Results of these simulations show that CareBot was able to perform some tasks such as fetching and carrying food, drinks, or medicine, and recognizing the absence of vital signs in order to alert caregivers (Madl and Franklin 2015)
Ethical decision-making model			Situationism (based on both deontological ethics and an empirical criterion)	This model was implemented in a virtual agent and tested with some hypothetical study cases. Results reported show how agent's actions can be influenced by emotional information (Cervantes et al. 2016)
MedEthEx			Situationism (based on both biomedical ethics and an empirical criterion)	MedEthEx was tested using case simulations. In particular, 173 American medical students divided into two group participated in the program assessment phase of this project. Results were compared overall taking into account final exam grades between the two groups and found no statistically significant difference (Fleetwood et al. 2000)
Ethical multiple agent system			Situationism (based on both legal reasoning and a degree of incoherence)	The algorithm proposed in this work was designed considering an uncooperative environment in which agents either compete to achieve their own goals, or collaborate, but not cooperate. This computational model was tested using formal demonstrations and theoretical cases (Cristani and Burato 2009)

Conclusion

As described in this paper, human beings are looking to delegate part of their decision-making power to AAs, thus increasing the scope of their activities. Endowing AAs with ethical mechanisms has been an initial approach in the field of machine ethics that attempts to deal with some likely issues that may arise in the relationship between AAs and human beings and coexist in a safe way, in harmony, and with confidence. This paper presented an analysis of the current status of ethical artificial agents. As a result of this review, a new taxonomy was proposed that aims to be useful for understanding advantages and limitations of AMAs based on the design strategy (i.e., top-down, bottom-up, hybrid, and implicit) and criteria (i.e., normative ethics, situationism, empirical, and non-malicious code) employed to make ethical decisions.

From the review presented, it is evident that currently there are no general artificial intelligence systems capable of making sophisticated moral decisions as humans do. However, there are a few prototypes that deal with certain moral issues, but these prototypes are still in their early stages of development. All these prototypes have been tested using basic cases where the test environment is well-defined and controlled, in contrast to real scenarios where critical or uncertain moral situations could arise unexpectedly. Moreover, the review illustrates that there is a long way to go (from a technological perspective) before this type of artificial agent can replace human judgment in difficult, surprising or ambiguous moral situations.

This paper concludes that because of the current interaction between AAs systems and humans, ethical mechanisms for AMAs are not a potential option for future autonomous systems, but an actual need to deal with. Perhaps their ethics could be different from human ethics, but at this moment, models of human ethics are the guides most used by researchers to develop AMAs.

References

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.

Alaieri, F., & Vellino, A. (2016). Ethical decision making in robots: Autonomy, trust and responsibility. In *International conference on social robotics* (pp. 159–168). Cham: Springer.

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology, 7*(3), 149–155.

Amstutz, M. R. (2013). *International ethics: Concepts, theories, and cases in global politics*. New York: Rowman & Littlefield Publishers.

Anderson, M., & Anderson, S. L. (2007a). Machine ethics: Creating an ethical intelligent agent. *AI Magazine, 28*(4), 15–26.

Anderson, M., & Anderson, S. L. (2007b). The status of machine ethics: A report from the AAAI symposium. *Minds and Machines, 17*(1), 1–10.

Anderson, M., & Anderson, S. L. (2008). Ethical healthcare agents. In M. Sordo, S. Vaidya, & L. C. Jain (Eds.), *Advanced computational intelligence paradigms in healthcare-3* (pp. 233–257). Berlin: Springer.

Anderson, M., & Anderson, S. L. (2010). Robot be good. *Scientific American, 303*(4), 72–77.

Anderson, M., & Anderson, S. L. (2014). Geneth: A general ethical dilemma analyzer. In *Twenty-eighth AAAI conference on artificial intelligence* (pp. 253–261).

Anderson, M., Anderson, S. L., & Armen, C. (2004). Towards machine ethics. In *Proceedings of the AOTP'04—The AAAI-04 workshop on agent organizations: Theory and practice*.

Anderson, M., Anderson, S. L., & Armen, C. (2005). Medethex: Toward a medical ethics advisor. In *Proceedings of the AAAI 2005 Fall symposium on caring machines: AI in elder care* (pp. 9–16).

Anderson, M., Anderson, S. L., & Armen, C. (2006a). An approach to computing ethics. *IEEE Intelligent Systems, 21*(4), 56–63.

Anderson, M., Anderson, S. L., & Armen, C. (2006b). Medethex: A prototype medical ethics advisor. In *Proceedings of the national conference on artificial intelligence* (Vol. 21, No. 2, pp. 1759–1765). Menlo Park, CA/Cambridge, MA: AAAI Press/MIT Press.

Andino, C. (2015). Place of ethics between technical knowledge. A philosophical approach. *Revista Científica de la UCSA, 2*(2), 85–94.

Arkin, R. (2009). *Governing lethal behavior in autonomous robots*. London: Chapman and Hall/CRC.

Arkin, R. C. (2010). The case for ethical autonomy in unmanned systems. *Journal of Military Ethics, 9*(4), 332–341.

Arkin, R. (2018). Lethal autonomous systems and the plight of the noncombatant. In R. Kiggins (Ed.), *The political economy of robots* (pp. 317–326). Cham: Springer.

Arkoudas, K., Bringsjord, S., & Bello, P. (2005). Toward ethical robots via mechanized deontic logic. In *AAAI Fall symposium on machine ethics* (pp. 17–23).

Ashrafian, H. (2015). Artificial intelligence and robot responsibilities: Innovating beyond rights. *Science and Engineering Ethics, 21*(2), 317–326.

Bandyopadhyay, D., & Sen, J. (2011). Internet of things: Applications and challenges in technology and standardization. *Wireless Personal Communications, 58*(1), 49–69.

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

Beauvisage, T. (2009). Computer usage in daily life. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 575–584). ACM.

Bedaf, S., Draper, H., Gelderblom, G. J., Sorell, T., & de Witte, L. (2016). Can a service robot which supports independent living of older people disobey a command? The views of older people, informal carers and professional caregivers on the acceptability of robots. *International Journal of Social Robotics, 8*(3), 409–420.

Belloni, A., Berger, A., Besson, V., Boissier, O., Bonnet, G., Bourgne, G., et al. (2014). Towards a framework to deal with ethical conflicts in autonomous agents and multi-agent systems. In *CEPE 2014 well-being, flourishing, and ICTs* (pp. 1–10).

Belloni, A., Berger, A., Boissier, O., Bonnet, G., Bourgne, G., Chardel, P. A., et al. (2015). Dealing with ethical conflicts in autonomous agents and multi-agent systems. In *1st International workshop on artificial intelligence and ethics at the 29th AAAI conference on artificial intelligence*.

Blass, J. A. (2016). Interactive learning and analogical chaining for moral and commonsense reasoning. In *Thirtieth AAAI conference on artificial intelligence* (pp. 4289–4290).

Blass, J. A., & Forbus, K. D. (2015). Moral decision-making by analogy: Generalizations versus exemplars. In *Twenty-ninth AAAI conference on artificial intelligence* (pp. 501–507).

Bonnemains, V., Saurel, C., & Tessier, C. (2018). Embedded ethics: Some technical and ethical challenges. *Ethics and Information Technology, 20*(1), 41–58.

Borenstein, J., & Arkin, R. (2019). Robots, ethics, and intimacy: The need for scientific research. In D. Berkich, & M. d'Alfonso (Eds.), *On the cognitive, ethical, and scientific dimensions of artificial intelligence* (pp. 299–309). Cham: Springer.

Borst, J. P., & Anderson, J. R. (2015). Using the ACT-R cognitive architecture in combination with fMRI data. In B. Forstmann, & E. J. Wagenmakers (Eds.), *An introduction to model-based cognitive neuroscience* (pp. 339–352). Berlin: Springer.

Brachman, R. J. (2002). Systems that know what they're doing. *IEEE Intelligent Systems, 17*(6), 67–71.

Briggs, G., & Scheutz, M. (2015). Sorry, I can't do that": Developing mechanisms to appropriately reject directives in human–robot interactions. In *2015 AAAI Fall symposium series* (pp. 1–5).

Bringsjord, S., Sundar, G. N., Thero, D., & Si, M. (2014). Akratic robots and the computational logic thereof. In *Proceedings of the IEEE 2014 international symposium on ethics in engineering, science, and technology* (pp. 1–8). IEEE Press.

Brundage, M. (2014). Limitations and risks of machine ethics. *Journal of Experimental & Theoretical Artificial Intelligence, 26*(3), 355–372.

Capraro, V., & Rand, D. G. (2018). Do the right thing: Experimental evidence that preferences for moral behavior, rather than equity or efficiency per se, drive human prosociality. *Forthcoming in Judgment and Decision Making, 13*(1), 99–111.

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.

Cervantes, J. A., Rosales, J. H., López, S., Ramos, F., & Ramos, M. (2017). Integrating a cognitive computational model of planning and decision-making considering affective information. *Cognitive Systems Research, 44,* 10–39.

Choi, D., & Langley, P. (2018). Evolution of the icarus cognitive architecture. *Cognitive Systems Research, 48,* 25–38.

Coeckelbergh, M. (2010). Moral appearances: Emotions, robots, and human morality. *Ethics and Information Technology, 12*(3), 235–241.

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. *Journal of Personality and Social Psychology, 104*(2), 216–235.

Cook, D. J., & Das, S. K. (2012). Pervasive computing at scale: Transforming the state of the art. *Pervasive and Mobile Computing, 8*(1), 22–35.

Cristani, M., & Burato, E. (2009). Approximate solutions of moral dilemmas in multiple agent system. *Knowledge and Information Systems, 18*(2), 157–181.

Czubenko, M., Kowalczuk, Z., & Ordys, A. (2015). Autonomous driver based on an intelligent system of decision-making. *Cognitive Computation, 7*(5), 569–581.

Dehghani, M., Tomai, E., Forbus, K. D., & Klenk, M. (2008). An integrated reasoning approach to moral decision-making. In *Twenty-third AAAI conference on artificial intelligence* (pp. 1280–1286).

Deng, B. (2015). Machine ethics: The robot's dilemma. *Nature, 523*(7558), 24–26.

Dennis, L. A., Fisher, M., Lincoln, N. K., Lisitsa, A., & Veres, S. M. (2016a). Practical verification of decision-making in agent-based autonomous systems. *Automated Software Engineering, 23*(3), 305–359.

Dennis, L., Fisher, M., Slavkovik, M., & Webster, M. (2016b). Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems, 77,* 1–14.

Epting, S. (2016). A different trolley problem: The limits of environmental justice and the promise of complex moral assessments for transportation infrastructure. *Science and Engineering Ethics, 22*(6), 1781–1795.

Erdur, M. (2018). Moral realism and the incompletability of morality. *The Journal of Value Inquiry, 52*(2), 227–237.

Fagin, R., Halpern, J. Y., & Vardi, M. Y. (1990). A nonstandard approach to the logical omniscience problem. In *Proceedings of the 3rd conference on theoretical aspects of reasoning about knowledge* (pp. 41–55). Morgan Kaufmann Publishers Inc.

Feil-Seifer, D., & Matarić, M. J. (2011). Socially assistive robotics. *IEEE Robotics and Automation Magazine, 18*(1), 24–31.

Ferrell, O. C., & Gresham, L. G. (1985). A contingency framework for understanding ethical decision making in marketing. *The Journal of Marketing, 49*(3), 87–96.

Fleetwood, J., Vaught, W., Feldman, D., Gracely, E., Kassutto, Z., & Novack, D. (2000). Medethex online: A computer-based learning program in medical ethics and communication skills. *Teaching and Learning in Medicine, 12*(2), 96–104.

Fumagalli, M., & Priori, A. (2012). Functional and clinical neuroanatomy of morality. *Brain, 135*(7), 2006–2021.

Gerdes, A., & Øhrstrøm, P. (2015). Issues in robot ethics seen through the lens of a moral turing test. *Journal of Information, Communication and Ethics in Society, 13*(2), 98–109.

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

Govindarajulu, N. S., Bringjsord, S., & Ghosh, R. (2018). *One formalization of virtue ethics via learning*. arXiv preprint arXiv:180507797.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107*(3), 1144–1154.

Greene, J., Rossi, F., Tasioulas, J., Venable, K. B., & Williams, B. C. (2016). Embedding ethical principles in collective decision support systems. In *Thirtieth AAAI conference on artificial intelligence* (pp. 4147–4151).

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*(5537), 2105–2108.

Guerini, M., Pianesi, F., & Stock, O. (2015). Is it morally acceptable for a system to lie to persuade me? In *Workshops at the twenty-ninth AAAI conference on artificial intelligence* (pp. 53–60).

Han, T. A., & Pereira, L. M. (2018). Evolutionary machine ethics. In O. Bendel (Ed.), *Handbuch Maschinenethik* (pp. 1–25). Wiesbaden: Springer.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron, 95*(2), 245–258.

Honarvar, A. R., & Ghasem-Aghaee, N. (2009). Casuist BDI-agent: A new extended BDI architecture with the capability of ethical reasoning. In *International conference on artificial intelligence and computational intelligence* (pp. 86–95). Berlin: Springer.

Howard, D., & Muntean, I. (2016). A minimalist model of the artificial autonomous moral agent (AAMA). In *The 2016 AAAI Spring symposium series* (pp. 217–225).

Hughes, G. J. (2001). *Routledge philosophy guidebook to Aristotle on ethics*. London: Routledge.

Kahn, P. H., Jr., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., et al. (2012). Do people hold a humanoid robot morally accountable for the harm it causes? In *Proceedings of the seventh annual ACM/IEEE international conference on Human–Robot Interaction* (pp. 33–40). ACM.

Kirchin, S. (Ed.). (2012). What is metaethics? In *Metaethics* (pp. 1–20). London: Palgrave Macmillan.

Kishi, T., Hashimoto, K., & Takanishi, A. (2017). Human like face and head mechanism. In A. Goswami, & P. Vadakkepat (Eds.), *Humanoid robotics: A reference* (pp. 1–26). Dordrecht: Springer.

Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberate judgments are based on common principles. *Psychological Review, 118*(1), 97–109.

Laird, J. E. (2008). Extending the soar cognitive architecture. *Frontiers in Artificial Intelligence and Applications, 171,* 224–235.

Laird, J. E., Kinkade, K. R., Mohan, S., & Xu, J. Z. (2012). Cognitive robotics using the soar cognitive architecture. In *Workshops at the twenty-sixth AAAI conference on artificial intelligence* (pp. 46–54).

Laird, J. E., Lebiere, C., & Rosenbloom, P. S. (2017). A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. *AI Magazine, 38*(4), 13–26.

Lombrozo, T. (2009). The role of moral commitments in moral judgment. *Cognitive Science, 33*(2), 273–286.

Long, L. N., & Kelley, T. D. (2010). Review of consciousness and the possibility of conscious robots. *Journal of Aerospace Computing, Information, and Communication, 7*(2), 68–84.

Madl, T., & Franklin, S. (2015). Constrained incrementalist moral decision making for a biologically inspired cognitive architecture. In R. Trappl (Ed.), *A construction manual for robots' ethical systems* (pp. 137–153). Cham: Springer.

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.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *Proceedings of the tenth annual ACM/IEEE international conference on human–robot interaction* (pp. 117–124). ACM.

Mermet, B., & Simon, G. (2016). Formal verification of ethical properties in multiagent systems. In *ECAI 2016 workshop on ethics in the design of intelligent agents (EDIA'16).* The Netherlands: The Hague.

Metta, G., Natale, L., Nori, F., Sandini, G., Vernon, D., Fadiga, L., et al. (2010). The iCub humanoid robot: An open-systems platform for research in cognitive development. *Neural Networks, 23*(8), 1125–1134.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences, 11*(4), 143–152.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems, 21*(4), 18–21.

Mordoch, E., Osterreicher, A., Guse, L., Roger, K., & Thompson, G. (2013). Use of social commitment robots in the care of elderly people with dementia: A literature review. *Maturitas, 74*(1), 14–20.

Mostafa, S. A., Ahmad, M. S., & Mustapha, A. (2019). Adjustable autonomy: A systematic literature review. *Artificial Intelligence Review, 51*(2), 149–186.

Mostafa, S. A., Mustapha, A., Mohammed, M. A., Ahmad, M. S., & Mahmoud, M. A. (2018). A fuzzy logic control in adjustable autonomy of a multi-agent system for an automated elderly movement monitoring application. *International Journal of Medical Informatics, 112,* 173–184.

Pellizzoni, S., Siegal, M., & Surian, L. (2010). The contact principle and utilitarian moral judgments in young children. *Developmental Science, 13*(2), 265–270.

Podschwadek, F. (2017). Do androids dream of normative endorsement? On the fallibility of artificial moral agents. *Artificial Intelligence and Law, 25*(3), 325–339.

Reig, S., Norman, S., Morales, C. G., Das, S., Steinfeld, A., & Forlizzi, J. (2018). A field study of pedestrians and autonomous vehicles. In *Proceedings of the 10th international conference on automotive user interfaces and interactive vehicular applications* (pp. 198–209). ACM.

Rodríguez, L. F., & Ramos, F. (2014). Development of computational models of emotions for autonomous agents: A review. *Cognitive Computation, 6*(3), 351–375.

Schaich Borg, J., Hynes, C., Van Horn, J., Grafton, S., & Sinnott-Armstrong, W. (2006). Consequences, action, and intention as factors in moral judgments: An fMRI investigation. *Journal of Cognitive Neuroscience, 18*(5), 803–817.

Scheutz, M., & Malle, B. F. (2014). Think and do the right thing: A plea for morally competent autonomous robots. In *Proceedings of the IEEE 2014 international symposium on ethics in engineering, science, and technology* (p. 9). IEEE Press.

Schroeder, M. (2017). Normative ethics and metaethics. In T. McPherson, & D. Plunkett (Eds.), *The Routledge handbook of metaethics* (pp. 674–686). London: Routledge.

Sharkey, A., & Sharkey, N. (2012). Granny and the robots: Ethical issues in robot care for the elderly. *Ethics and Information Technology, 14*(1), 27–40.

Shigemi, S. (2018). ASIMO and humanoid robot research at Honda. In A. Goswami, & P. Vadakkepat (Eds.), *Humanoid robotics: A reference* (pp. 1–36). Springer.

Tikhanoff, V., Cangelosi, A., & Metta, G. (2011). Integration of speech and action in humanoid robots: iCub simulation experiments. *IEEE Transactions on Autonomous Mental Development, 3*(1), 17–29.

Trafton, G., Hiatt, L., Harrison, A., Tamborello, F., Khemlani, S., & Schultz, A. (2013). ACT-R/E: An embodied cognitive architecture for human–robot interaction. *Journal of Human–Robot Interaction, 2*(1), 30–55.

Van Riemsdijk, M. B., Jonker, C.M., & Lesser, V. (2015). Creating socially adaptive electronic partners: Interaction, reasoning and ethical challenges. In *Proceedings of the 2015 international conference on autonomous agents and multiagent systems* (pp. 1201–1206). International Foundation for Autonomous Agents and Multiagent Systems.

Van Staveren, I. (2007). Beyond utilitarianism and deontology: Ethics in economics. *Review of Political Economy, 19*(1), 21–35.

Van Wynsberghe, A., & Robbins, S. (2018). Critiquing the reasons for making artificial moral agents. *Science and Engineering Ethics, 25*(3), 1–17.

Vanderelst, D., & Winfield, A. (2018). An architecture for ethical robots inspired by the simulation theory of cognition. *Cognitive Systems Research, 48,* 56–66.

Vernon, D., Metta, G., & Sandini, G. (2007). A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents. *IEEE Transactions on Evolutionary Computation, 11*(2), 151–180.

Viroli, M., Pianini, D., Montagna, S., & Stevenson, G. (2012). Pervasive ecosystems: A coordination model based on semantic chemistry. In *Proceedings of the 27th annual ACM symposium on applied computing* (pp. 295–302). ACM.

Von der Pfordten, D. (2012). Five elements of normative ethics—A general theory of normative individualism. *Ethical Theory and Moral Practice, 15*(4), 449–471.

Von Wright, G. H. (1951). Deontic logic. *Mind, 60*(237), 1–15.

Waldrop, M. M. (2015). Autonomous vehicles: No drivers required. *Nature News, 518*(7537), 20.

Walker, L. J., & Hennig, K. H. (2004). Differing conceptions of moral exemplarity: Just, brave, and caring. *Journal of Personality and Social Psychology, 86*(4), 629–647.

Wallach, W. (2008). Implementing moral decision making faculties in computers and robots. *AI & Society, 22*(4), 463–475.

Wallach, W. (2010). Robot minds and human ethics: The need for a comprehensive model of moral decision making. *Ethics and Information Technology, 12*(3), 243–250.

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.

Wang, S., Wan, J., Zhang, D., Li, D., & Zhang, C. (2016). Towards smart factory for industry 4.0: A self-organized multi-agent system with big data based feedback and coordination. *Computer Networks, 101,* 158–168.

Wellman, M. P., & Rajan, U. (2017). Ethical issues for autonomous trading agents. *Minds and Machines, 27*(4), 609–624.

Winfield, A. F., Blum, C., & Liu, W. (2014). Towards an ethical robot: Internal models, consequences and ethical action selection. In *Conference towards autonomous robotic systems* (pp. 85–96). Cham: Springer.

Yampolskiy, R. V. (2013). Artificial intelligence safety engineering: Why machine ethics is a wrong approach. In V. Müller (Ed.), *Philosophy and theory of artificial intelligence* (pp. 389–396). Berlin: Springer.

Young, L., & Durwin, A. (2013). Moral realism as moral motivation: The impact of meta-ethics on everyday decision-making. *Journal of Experimental Social Psychology, 49*(2), 302–306.

Zambonelli, F., & Viroli, M. (2011). A survey on nature-inspired metaphors for pervasive service ecosystems. *International Journal of Pervasive Computing and Communications, 7*(3), 186–204.

Zieba, S., Polet, P., Vanderhaegen, F., & Debernard, S. (2010). Principles of adjustable autonomy: A framework for resilient human–machine cooperation. *Cognition, Technology & Work, 12*(3), 193–203.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

José-Antonio Cervantes[1] · Sonia López[1] · Luis-Felipe Rodríguez[2] · Salvador Cervantes[1] · Francisco Cervantes[3] · Félix Ramos[4]

Sonia López
sonia.lopez@valles.udg.mx

Luis-Felipe Rodríguez
luis.rodriguez@itson.edu.mx

Salvador Cervantes
salvador.cervantes@valles.udg.mx

Francisco Cervantes
fcervantes@iteso.mx

Félix Ramos
framos@gdl.cinvestav.mx

[1] Department of Computer Science and Engineering, Centro Universitario de los Valles, Universidad de Guadalajara, Carretera Guadalajara – Ameca Km. 45.5, 46600 Ameca, Mexico

[2] Department of Computer Science, Instituto Tecnológico de Sonora, Sonora, Mexico

[3] Department of Electronics, Systems and Informatics, Instituto Tecnológico y de Estudios Superiores de Occidente, Tlaquepaque, Mexico

[4] Department of Computer Science, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, Guadalajara, Mexico