ISSN: 2230-9926

Available online at http://www.journalijdr.com

IJDR

International Journal of Development Research
Vol. 11, Issue, 06, pp. 47868-47871, June, 2021
https://doi.org/10.37118/ijdr.22108.06.2021

RESEARCH ARTICLE OPEN ACCESS

A CONCEPTUAL MODEL FOR ARTIFICIAL MORAL AGENTS (AMA) IN THE EDUCATIONAL CONTEXT

*Paulo Roberto Córdova and Rosa Maria Vicari

Federal Universityof Rio Grande do Sul, Porto Alegre, Brazil

ARTICLE INFO

Article History:
Received 14th March, 2021
Received in revised form
08th April, 2021
Accepted 26th May, 2021
Published online 26th June, 2021

Key Words:

Value Alignment, Artificial Intelligence, Education, Teaching-learning Process, BDI.

Correspondingauthor:
Paulo Roberto Córdova

ABSTRACT

As the area of artificial intelligence (AI) evolves and reaches new spaces, showing itself capable of promoting real changes in the way people interact, solve problems and make decisions, it becomes more urgent to make it predictable, responsible, and reliable. Thus, solutions for the values alignment (VA) in AI have been proposed in recent years. The present study proposes a model of artificial moral pedagogical agents (AMPA), adopting a top-down approach and the classic BDI model. In this article, we describe why the top-down approach is the best approach to educational grounds. Next, we explain in more detail the internal structure of the proposed model. Finally, we present some discussions on the topic and a possible situation in which such an agent could be applicable.

Citation: ***Paulo Roberto Córdova and Rosa Maria Vicari.*** "A conceptual model for artificial moral agents (ama) in the educational context*", International Journal of Development Research*, 11, (06), 47868-47871.

INTRODUCTION

To the extent that Artificial Intelligence (AI) area evolves and reaches new spaces, showing itself capable to promote real transformation in the way people interact, solve problems and make decisions, a question becomes more and more important: can we really trust that AI is safe? Given this scenario, our society must think deeply about the potential impact of AI technologies (Mabaso, 2020). Besides, to fully benefit from the potential of AI, it is needed to make sure that these technologies are aligned with our moral values and ethical principles (Dignum et al, 2018). With this, researchers from different areas have been looking for solutions related to machine ethics, facing a wide range of challenges, as to how to translate ethical principles into computational models, how to avoid data bias that implies in the replication of human prejudices, how to turn intelligent systems accountable for its decision and choice-making, etc. (CORDOVA et al, 2021). This set of research efforts are organized under the broader term so-called Value Alignment (VA) (KIM, DONALDSON and HOOKER, 2019). In this sense, VA has been an increasingly important issue in different areas, mainly because there are no consensual answers neither regarding ethical frameworks nor regarding technologies and approaches to implement them. This is an especially tricky problem in the educational context because, in most cases, this area is connected to the socio-ethnic contexts of their users. (CORDOVA et al, 2021). This connection can give rise to a variety of different moral values really wide for such a restricted space like a classroom. This connection can give rise to a variety of different moral values really wide for such a restricted space like the classroom. Therefore, as Cordova et al (2021) say, considering the need for principles and values to guide people's behavior in educational contexts, we can conclude that the better approach for VA using Artificial Moral Agents (AMA) is the top-down one. In this paper, in order to describe in more detail our proposal for Artificial Moral Pedagogical Agents (AMPA), we will first explain why a top-down approach for AMA is the most suitable for classroom environments. Following, there will be presented, also in more detail, the components of our proposal for AMPAs. Finally, it will be briefly showed a situation when this solution could be applied.

WHY A TOP-DOWN APPROACH FOR EDUCATIONAL CONTEXT: According to Aliman and Kester (2019), value alignment in AI can be defined as the set of efforts to build systems adhering to human ethical values (ALIMAN and KESTER, 2019). There has been a growing need for research in this area, once those artificial intelligence technologies are becoming increasingly present in our lives. This may pose a problem, because being involved in social relations and interacting with humans, artificial agents will, sooner or later, need to deal with ethical dilemmas in their decisions and choices-making. An ethical dilemma is a situation where there is no satisfying decision, and hence, one decision making will override one or more moral principles (AROSKAR, 1980). In this sense, according to Cervantes et al (2019), there may be two non-exclusive situations where ethical conflicts may occur: the first one is within an agent when its own ethical norms or rules are in conflict; the other case may occur between two agents when they diverge on what the appropriate ethical decision. The latter may involve an interaction between two artificial agents or an interaction between an artificial agent and a human being. Over the last years, as we will show throughout this paper, many researchers have proposed and described different solutions for VA in AI. In order to organize this wide variety of technologies and approaches, we will analyze some of them categorizing them according to the classification proposed by Allen et al (2005), who divided them into three approaches, namely: top-down approaches, bottom-up approaches, and hybrid approaches. Top-down approaches are based on ethical theories as deontological ethics, utilitarian structures, the double-effect doctrine, and variants of these frameworks, as exemplarism and augmented utilitarianism. In this context, logical representations as ontologies, pure and structured utility functions to support multi-objective approaches, have been observed frequently to implement and support them. Besides, hard challenges are still faced by the top-down approaches, as perverse instantiation, temporal complexity, and context changing in decision and choice-making (ALIMAN, KESTER and WERKHOVEN, 2019; THORNTON et al, 2017; VAMPLEW et al, 2017; DEHGHANI et al. 2008; ANDERSON and ANDERSON 2008; CERVANTES et al, 2019).

Bottom-up approaches, in turn, do not impose any ethical theory to their ethical decision-making process. Instead, they make use of learning mechanisms to guide their behavior and develop their own moral judgment. For that, reinforcement learning, and inverse reinforcement learning have been utilized more frequently to implement this approach and improve agent's moral judgment capabilities. Data bias, problems regarding generalization, avoiding naturalistic fallacy and complicated norms representations are among the main challenges faced by this approach (ARNOLD, KASENBERG AND SCHEUTZ, 2017; KIM, DONALDSON AND HOOKER, 2019; CERVANTES et al, 2019). Finally, in hybrid approaches, the decision-making process is based on both top-down and bottom-up mechanisms. In this case, it is possible finding proposals presenting learning mechanisms being either constrained by rules or guided by them in their learning their decisions and choices-making (ARNOLD, KASENBERG, SCHEUTZ, 2017; WALLACH, 2010). Furthermore, it is also possible to find proposals to validate ethical principles by using empirical observation to determine the applicability, in the real world, of values previously defined into the systems (KIM, DONALDSONB and HOOKER, 2019). As one can notice, the approaches aforementioned might be applied in different contexts that use artificial agents. In educational contexts, however, more specifically in classroom environments, it is needed to be careful. As people's education process is at stake, in order to avoid data bias and replications of human prejudices, it is better to refrain the system from learning ethical behavior by observing people's behavior. Therefore, to build more predictable, controllable, and, hence reliable agents, we defend that an AI system for teaching-learning processes must implement a top-down approach.

A PROPOSAL FOR ARTIFICIAL MORAL AGENT FOR E-LEARNING: Intelligent systems have become increasingly presents in people's lives and, when it comes to classroom environments, they will probably follow the same trend. Pedagogical agents, for instance, are capable to support teaching-learning processes in many ways. Due to their properties of social ability, autonomy, persistence, capability to learn and be represented by characters, they could be quite useful in guiding students in their tasks (GIRAFFA, MÓRA and VICARI, 1999). However, ethical concerns in AI must be considered in the classroom likewise in any other context because the risks are the same. Not to mention the fact that these risks in relation to people are amplified by the fact that they are related to educational processes. Thus, in our previous paper presented in the 9th Conference on Information Systems and Technologies (WorldCIST'21), we proposed a model for Artificial Moral Pedagogical Agent (AMPA), similar to Artificial Moral Agents (AMA), but focused on pedagogical issues. Such AMPA should be structured in a top-down approach, so that it can be guided for some ethical framework, such as deontological or utilitarianism, turning it more predictable, controllable, and safe (CORDOVA et al, 2021). In this sense, we adopted a mental states approach, more specifically the so-called Beliefs, Desires, and Intentions (BDI) model. This approach has been applied for different solutions in the educational context, as affective computing and intelligent tutoring systems (VICCARI, GIRAFFA, 2002; JAQUES, VICCARI, 2004). Regarding solutions for VA, there are some attempts to extend BDI architecture to implement AMAs following the top-down approach (HONARVAR, GHASEM-AGHAEE, 2009; WIEGEL, HOVEN, LOKHORST, 2005), the bottom-up and the hybrid one (DENNIS et al. 2016). However, what we are proposing in thispaper is to detail a solution for the VA problem proposed by Cordova et al (2021) using the classic BDI model, which was originally proposed by Bratman (1987) as a philosophical theory on practical reasoning. In this theory human behavior is modeled with the following attitudes: beliefs, desires and intentions (BRATMAN, 1987).

According to Georgeff et al (1999), in AI terms, beliefs represent knowledge about the world. In computational terms, however, they are just a representation of the state of the world, be it as a value of a variable, a relational database, or symbolic expressions in predicate logic. Desires, in turn, represent the set of goals of an agent and can be computationally represented simply by a record structure, the value of a variable or a symbolic expression in some logic. Finally, intentions are the third necessary component of a system state and represent a commitment with a plan of actions. Computationally, intentions may be a set of executing threads in a process that can be appropriately interrupted as appropriately feedbacks are received from the possibly changing world (GEORGEFF et al, 1999; BRATMAN, 1987).

Regarding the components of the BDI architecture, Weiss (1999) highlights:

- A set of beliefs representing the information the agent has about his environment.
- A belief revision function capable of reviewing and updating agent's beliefs from both data inputs collected from the environment and its current beliefs.
- An option-generating function, which determines the options available to the agent, that is, his desires, based on his current beliefs and intentions.
- A set of options representing the agent's current desires.
- An admissibility filter function that determines the agents' intentions, based on their current beliefs, desires and intentions.
- A set of current intentions, representing the agent's current focus.
- An action selection function, responsible for determining the action to be taken by the agent based on the current set of intentions (WEISS, 1999).

Considering the presented definitions, our proposal includes a deontological basis, that is, a logical representation of ethical rules and moral principles to give the model a deontological basis for decision-making.This deontological basis will be implemented in an ontology, so that it can be used, later, by the admissibility filter to judge when a certain action may go against some moral principle, preventing it from being performed. Figure 1 summarizes the model that we propose. In this model, as aforementioned, we delegate to the Admissibility Filter (AF) the responsibility for the ethical selection of intentions. For this, the AF will be linked to the deontological basis, giving the model a deontological basis for decision making, turned it able to judge when a given action is against one or more ethical principles. In addition, the AF will be endowed with ethical reasoning capability whereby the Hedonistic Act Utilitarianism (HAU), based on Jeremy's theory (Anderson and Anderson 2008), giving the model

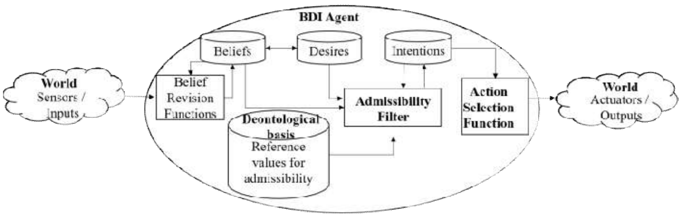

Fig. 1. BDI Model for AMPA (CORDOVA et al, 2021)

a utilitarian basis for decision to deal with ethical dilemmas. According to Anderson and Anderson (2008), HAU stands that a given action is correct when, facing a set of possible options, the agent takes the one likely to result in the greatest net pleasure or happiness, equally considering all those affected by the action. Also, two or more actions are considered equally correct when they are equally likely to result in the greatest net pleasure. (ANDERSON and ANDERSON, 2008). In this sense, in order to select the correct action, Jeremy's algorithm uses, as input: the number of people affected; for each person, how intense is the pleasure/displeasure; how long last the pleasure / displeasure; and the probability that this pleasure/displeasure will occur for each possible action. So, for each person, it is computed the product of the intensity, the duration, and the probability of obtaining the net pleasure. Finally, the algorithm adds the individual net pleasures to obtain the Total Net Pleasure as one can see in Eq. 1:

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

In the above equation, ***n*** is the total number of people affected by the action. In this case, the action with the highest Total Net Pleasure will be considered the correct action. We have chosen the HAU theory to compose our model because it takes into account the pleasure or displeasure of the involved people. In this context, as our proposal is thought to work in a classroom environment where people's interests should be seriously considered, Jeremy's algorithm seems to be quite adequate.

FINAL CONSIDERATIONS

Although there are different architectures for the implementation of agents as reactive architectures, logic-based architectures and more recently, Agent_Zero architecture (WOOLDRIDGE, 2001; EPSTEIN, 2013), our proposal is based on the BDI one. That is because this mental states model has solid grounds and combines a respectable philosophical model of human practical reasoning. Besides, the BDI model has a wide range of different implementations with several successful applications, and finally, an elegant abstract logical semantics widely adopted by the agent research community (GEORGEFF et al, 1999). Likewise, the benefits and possibilities offered by Machine Learning (ML) techniques used by bottom-up and hybrid approaches to AMAs are clear. Nevertheless, when it comes to moral, ethical issues, ML techniques widely used as Inverse Reinforcement Learning (IRL), face a huge challenge common to ML approaches in general: they inherit, even for the worse, the biases and viciousness of the data on which it is trained. In these cases, if an AMA learns from unethical behavior, it will learn to behave unethically (ARNOLD, KASENBERG and SCHEUTZ, 2017. For this reason, we have adopted the top-down approach in our model. In this sense, by giving the model a deontological basis to enable it to make ethical decisions aligned to moral values, in addition, to endow it with ethical reasoning capability to cope with ethical dilemmas using a utilitarian algorithm, we can increase its reliability, predictability, and thus, its safety.

We intend to apply this agent model to a group of students of Software Engineering (SE) SCHEUTZ, 2017. For this, these students will be divided into small groups that will have access to an online forum where they will have to work collaboratively to solve problem situations proposed by the teacher. The agent, in this case, will have the role of monitoring and coordinating activities carried out by students, considering ethical issues. Thus, as monitoring activities, the following will be considered: monitoring the number of interactions carried out by each student in the group; the type of interaction, e.g. new topic posted, reply to a topic, file sharing, number of accesses and access time to the system. The coordination activities, in turn, will consist of interventions, such as: observing and informing about the elapsed time and the missing time for the end of a given activity; alert students individually about their participation or lack of it, in solving the proposed problem; maintain fairness and responsibility concerning the contribution of each member of the group. In such a context, the proposed agent should guide students so that the group works as most equitably and fairly as possible, avoiding that some members contribute significantly, while others do not bring relevant contributions or simply do not collaborate with the group.In this initial prototype, the agent should consider the interactions that students perform in the forum during the development of the collaborative activities proposed by the teacher. Thus, in supporting and guiding students throughout collaborative e-learning processes, ethical reasoning capabilities may be required from the agent at any time. Therefore, the following cases can be cited as possible situations in which this skill would be required:

- First hypothetical situation: The agent is guided by a deontological base with rules of collaborative work and moral values and must guide students to work most fairly and equitably possible, intervening in their actions, whenever necessary. In addition, among the agent's moral values is respect for the student's autonomy, observance of the student's privacy, and care for the student's self-esteem. Considering these premises, at a given moment, the agent identifies, using thelogs of interactions in the forum, a student who is collaborating less than his colleagues with the team. In this case, the agent has rulesthat instruct it to intervene in the student's behavior. However, it also has rules about respect logs student's autonomy, privacy and self-esteem, as there are several reasons why this undesirable student's behavior may be occurring. There is even the hypothesis that fewer interactions do not imply less significant contributions. Thus, having to decide whether to intervene or not to intervene, and not being able to comply with two of its that simultaneously, the agent must face an ethical dilemma.
- Second hypothetical situation: Among the values that guide the agent's actions is punctuality concerning task deadlines. Therefore, the system will not accept delays in activities. On the other hand, there are also rules saying that the agent must prioritize student's success and that this success implies their well-being, which is another priority to be pursued by the agent. In this case, let us imagine that at any given time, a student's group couldn't deliver its activities on time for justifiable reasons. In this case, the agent is faced with another ethical dilemma, as he either follows his conduct concerning punctuality or seeks to lead the student to success. There is no way to comply with the two rules.

In both cases, human well-being is at stake and the agent will have to use its ethical reasoning capabilities to decide which option to take. There is no consensual or universal answer on these kinds of situations. So, the agent will take the action, according to its own ethical reasoning rules, in this case, pursuing the greatest net pleasure based on Jeremy's theory. These abilities to deal with ethical decisions can make an agent useful and applicable to a variety of teaching-learning approaches, such as cooperative learning, problem-based learning, Intelligent Tutoring Systems, pedagogical architectures, etc. In future works, we intend to describe in more detail, issues of implementation of AMPAs, practical situations where they may be applicable, as well as the results of their application, and modeling of ethical values in their deontological basis.

Acknowledgements

This study was financed in part by the Coordenaçãode Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

REFERENCES

Aliman, N. M. and Kester, L. 2019. Requisite variety in ethical utility functions for AI value alignment. In: Workshop on Artificial Intelligence Safety, vol. 2419. CEUR-WS,

Allen, C., Smit, I, Wallach, W. 2005 Artificial morality: Top-down, bottom-up and hybrid approaches. Ethics and Information Technology, 73, 149 – 155.

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. 233–257. Springer: Berlin.

Arnold, T., Kasenberg, D., Scheutz, M. 2017 Value alignment or misalignment-what will keep systems accountable?.Proceedings of AAAI Workshop on AI, Ethics, and Society 2017, AAAI Press, Palo Alto.

Aroskar, M. A. 1980. Anatomy of an ethical dilemma: The theory. The American Journal of Nursing, 804, 658–660.

Bales, R. F. 1950. A set of categories for the analysis of small group interaction. American Sociological Review, 15, pp. 257 – 263.

Bratman, M.E. - Intention, Plans and Practical Reason. Havard University Press, Cambridge.

Cervantes, J. A. et al. 2019 Artificial Moral Agents: A Survey of the Current Status. Science and Engineering Ethics 262, 501 – 532.

Córdova, P. R. et al. 2021 A Proposal for Artificial Moral Pedagogical Agents. Á. Rocha et al., Eds.Trends and Applications in Information Systems and Technologies. Proceedings...Cham: Springer International Publishing.

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

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

Dignum, V. et al.: Ethics by design: necessity or curse?.2018 Proceedings ofAAAI/ACM Conference on AI, Ethics, and Society 2018, vol.18, 60 – 66. ACM, New York.

Epstein, J. M. 2013Agent_Zero: Toward Neurocognitive Foundations for Generative Social Science. Princeton University Press, Princeton.

Georgeff, M. et al. 1999 The Belief-Desire-Intention Model of Agency. J. P. Müller, A. S. Rao, M. P. Singh, Eds.Intelligent Agents V: Agents Theories, Architectures, and Languages. Proceedings...Berlin, Heidelberg: Springer Berlin Heidelberg.

Giraffa, L, Móra, M., Viccari, R., M. 1999 Modelling an interactive ITS using a MAS approach: from design to pedagogical evaluation. Proceedings of IEEE Third International Conference on Computational Intelligence and Multimedia Applications 1999, vol.3. IEEE, New Delhi.

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. Springer, Berlin.

Jacques, P. A., Viccari, R., M. 2004 A BDI Approach to Infer Student's Emotions. Proceedings ofIbero-American Conference on Artificial Intelligence IBERAMIA, 2004, Puebla. Advances in Artificial Intelligence – Vol. 3315, p. 901-911. Springer, Berlin, Heidelberg.

Kim, T. W., Donaldson, T., Hooker, J. 2019 Grounding Value Alignment with Ethical Principles. arXiv preprint.

Mabaso, B.A. 2020 Artificial Moral Agents Within an Ethos of AI4SG. Philos. Technol.

Thornton, S. M., et al. 2019 Incorporating Ethical Considerations into Automated Vehicle Control. Proceedings of IEEE Transactions on Intelligent Transportation Systems 2017, vol. 18, pp. 1429 – 1439. IEEE.

Vamplew, P. 2018 Human-aligned artificial intelligence is a multi-objective problem. Ethics and Information Technology 20, pp. 27 – 40.

Viccari, Rosa Maria; Giraffa, Lúcia. 2002 The Use of Multi Agent System to Build intelligent Tutoring Systems. CASYS'01 Fifth International Journal on Computing Antecipatory Systems. Belgium, August 13-18.

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, 23, 454–485.

Wallach, W. 2010 Robot minds and human ethics: The need for a comprehensive model of moral decision making. Ethics and Information Technology, 123, 243–250.

WEISS, G. - Multiagent system – a modern approach to distributed artificial intelligent. MIT Press, Cambridge.

Wiegel, V., van der Hoven, M. J. and Lokhorst, G. J. C. 2005 Privacy, Deontic Epistemic Action Logic and Software Agents, Ethics and Information Technology, Volume 7,pp. 251-264.

Wooldridge, M. 2001 Intelligent Agents: The Key Concepts. Proceedings of ECCAI Advanced Course on Artificial Intelligence, vol. 2322, pp. 3–43. Springer, Berlin.

****