LUPINE PUBLISHERS

ISSN: 2643-6736

Advances in Robotics & Mechanical Engineering

DOI: 10.32474/ARME.2019.02.000136

Mini Review

The Need for Ethical Artificial Intelligence and the Question of Embedding Moral and Ethical Principles

Thibault de Swarte*

Department of Sponsored Research and Consultancy Division, IMT Atlantique, France

***Corresponding author:** Thibault de Swarte, Department of Sponsored Research and Consultancy Division, IMT Atlantique, France

Received: December 12, 2019 **Published:** December 16, 2019

Introduction

The issue of Facebook moderators made topical headlines in 2019. When confronted with horrifying images such as terrorist assaults filmed online, moderators are required to be extremely reactive in order to eliminate the scenes violating human dignity from the social network as rapidly as possible. In [1], my colleagues and I explored whether it was possible or not to model human values such as virtue.Modeling could eventually make it possible to automatize all or part of the moderators' arduous and ungrateful work. In a first part, this article deals with the need for a reflection on ethical Artificial Intelligence (AI). After providing a definition of what AI is, I then discuss how ethical rules could be implemented in a system using AI. In a second part, I ask myself whether it is possible to embed moral and ethical principles. Using a utility function can help agents to make ethical decisions, but it is also important to be cautious about the limitations of such a sometimes-simplistic approach. Ethics must be appreciated in the social and technical context in which it is first developed and then implemented [2].

The Need for Ethical Artificial Intelligence

Artificial intelligence

A definition AI can be described by a universal triad, namely the data brought by the environment, the operations defined as the logic which mimic human behavior and finally, a control phase aiming at retroacting over its previous actions. Its definition is essentially based on two complementary views: one focused on the behavior and how it acts, especially as a human and the other which emphasizes the reasoning processes and how it reproduces human skills [3]. However, both points of view insist on the rational behavior that an AI must have. Moreover, it is important to pay attention to which kind of AI we are dealing with: strong or weak AI [4]. Weak AI, also known as narrow AI is shaped by behaviors answering to observable and specific tasks that may be represented by a decisional tree. On the other side, the strong AI, or artificial general intelligence, can copy human-like mental states. For this type of AI, this means that decision abilities or ethical behavior are issues that need to be taken care of. Finally a strong AI could find the closest solution to the given objective and learn with an external reversions. The latter is the one that is posing unprecedented problems that researchers are just starting to study. In fact, a system embedding strong AI is able to learn without human assistance or injection of additional data since the AI algorithm generates its own knowledge. Therefore, the exterior observer or the user of such agents will no longer know what the AI knows, what it is capable of doing nor the decisions it is going to take. Hence the need to establish an ethical framework that defines an area of action and prevents the system from taking decisions contrary to the ethics.

How to implement ethics rules ?

In order to implement ethical rules, there are two approaches. The first one named top down approach is based on ethical rule-abiding machines [5-6]. The strategy is to respect unconditionally the ethical principles related to morality such as "Do not kill". However, without understanding the potential consequences of the empirical decisions taken, an AI system creates numerous approximations that are a significant drawback. This can make rules conflict even for the 3 laws of robotics [7]; it may also lead to unintended consequences due to added rules [8]. It should be noted that even inaction can be taken into account for injuring humans. Moreover, the complexity of interaction between humans' priorities may lead to interpersonal inappropriate comparisons of various added laws [9]. The second method called bottom up, focuses on case studies in order to learn general concepts. The case studies make it possible for a strong AI to autonomously learn wrong and biased principles and even generalize them by applying them in new situations it encounters. These types of approaches are considered to be dangerous. In fact, in this process of learning, basic ethical concepts are acquired through a comprehensive assessment of the environment and its compliance with previous knowledge. This is done without any bottom-up procedure. The result of this learning will be taken into account for future decision making [10,11]. Eventually, in the case where an AI algorithm is facing a new

situation it has not encountered before, the extrapolation without a control phase may result in perilous situations for humans [12].

Artificial Intelligence Embedding Moral and Ethical Principles

A utilitarian approach of ethics consists in choosing in the case of a set of possibilities the solution that leads to the action consequently maximizing intrinsic good or net pleasure [13]. This involves quantifying Good or Evilfrom a given situation. However, certain situations supported by ethical reasons with an empirical study may prohibit the combined execution of certain actions. These complex cases are at the origin of dilemmas and the ethical principles do not make it possible to establish a preference. Therefore, autonomous agents need to be endowed with the ability to distinguish the most desirable option in the light of the ethical principles involved.

To achieve this goal, this article proposes in the following subsections a method called a utility function as a mean of avoiding ethical dilemmas.

Using a utility function to help agents make ethical decisions

In order to achieve this goal, a number of solutions have been proposed [14]. One of them it the utility function, also known as objective function. This function is used to assign values to outcomes or decisions. The optimal solution is the one that maximizes the utility function.This approach based on quantitative ethics determines which action maximizes benefit and minimizes harm. Its objective is to make it possible for an AI algorithm to take the right decisions particularly when it encounters an ethical dilemma. From a mathematical point of view, the utility function takes a state or a situation as an input parameter and gives as a result an output which is a number [15]. This number is an indication of how good the given state or situation is for the agent. The agent should then make the decision that leads to the state that maximizes the utility function. For instance, let's take the case of an autonomous vehicle, and let's assume that the car is in a situation where harm is unavoidable, and that it would inevitably hit either two men on the road or crash into a wall killing the passenger it is carrying [16]. Based on our previous utilitarian ethics definition, the decision that will minimize harm is the one that will lead to kill as few people as possible. Therefore, the car should crash and kill the passenger to save the two pedestrians because the utility function of this outcome is the highest. The same reasoning applies for military drones when they have to choose between multiple outcomes that involve moral and ethical principles.

Autonomous cars embedding AI algorithms using the utility function are not yet marketed. Some models available to the general public have an autopilot mode that still requires the presence of a human being behind the steering wheel who will make a decision in case of a problem. Fully autonomous cars still ride in test environments [17]. In the near future, the people who are likely to buy this type of car will primarily be public institutions such as municipalities. For instance, the city of Helsinki is testing an autonomous bus line, RoboBusLine which carries passengers on a defined road with a limited speed and an autonomous shuttle is also in service in Las Vegas [18]. However, these are still prototypes in a test phase with an operator on board. The other customers that may be interested in using autonomous vehicles are companies that make deliveries given the advantage of automating the tasks resulting in cost reduction and efficiency. In fact, Amazon, FedEx and UPS are investigating solutions for driverless trucks. The utility function is currently under investigation as an active solution to avoid ethical dilemmas non-modifying on-policy [19]. Autonomous robots are expanding, and the aim is not only to deal with ethical dilemmas but also to reduce uncertainty by quantifying problems such as exploration or unknown mapping; both can be stochastically defined (Shannon or Rényi's entropy) [20,21]. Describing and taking actions in a world incompletely defined can be done with the help of estimators but utility functions describe the perceptual state in line with the rules, and an active strategy can hence be implemented. This is already done for robot vision for example [22].

Limits and dangers of the utilitarian approach

The approach previously described and consisting in quantifying situations and assessing them with a utility function through a model has its own limits as far as strong AI is concerned. For weak AI, engineers at the design stage can implement decision trees to establish rules. They can anticipate the behavior of the AI more easily. On the other hand, as mentioned in section 2.1, advanced AI systems learn directly from the environment and adapt accordingly. By doing so, an external observer cannot always predict or anticipate the actions of such systems [23]. This is true for the Alpha Go algorithm that is taking decisions and implements strategies even experts in the game cannot understand although it leads to an optimum solution. The intelligent agent behaves like a black box whose internal functioning is unknown. This is particularly dangerous when it comes to autonomous vehicles or UAV drones that put human life at stake. Using only a utility function to decide whether or not a UAV could be used in an armed conflict could be considered as a war crime. Indeed, it is essential to test AI algorithms in different environments [23] and cover as many situations as possible before they are registered for use. This involves confronting algorithms with different situations and ensuring they behave properly by taking the most ethical decisions possible. It will then be possible to identify anomalies and correct them immediately.

Conclusion

In just a few years, artificial intelligence has become a strategic issue in Europe, the USA and China. For fundamental considerations of balance of powers between that of the GAFAM on the one hand and ethics at the service of the greatest number of people on the other hand, it will be crucial to develop an "ethics in context". This is more than computational ethics and utility functions developed by economists. It will be possible to implement an embedded code of ethics using artificial ethical agents, but only if ethical principles remain submitted to a democratic deliberation involving all

participants.My future research will focus on the question of "human values and AI" at different levels: universality, continentality (USA vs. China for example), nation-state and local communities using AI.

References

1. Vallée T, Bonnet G, de Swarte T (2018) Modélisation de valeurs humaines: le cas des vertus dans les jeux hédoniques. Revue d'Intelligence Artificielle 32(4): 519.
2. De Swarte T, Boufous, O Escalle P (2019) Artificial intelligence, ethics and human values: the cases of military drones and companion robots. Artificial Life and Robotics 24(3): 291-296.
3. Russell SJ, Norvig P (1995) Artificial Intelligence: A Modern Approach pp. 4-5.
4. Bringsjord S, Schimanski B (2003) What is Artificial Intelligence? Psychometric AI as an Answer p. 6
5. M Powers T (2006) Prospects for a Kantian Machine pp. 48-50.
6. Hanson R (2009) Prefer Law to Values.
7. Asimov I (1950) "Runaround". I Robot pp. 40.
8. Philip P (2003) Akrasia Collective and Individual.
9. Wallach W, Allen C, Smit I (2008) Machine Morality: Bottom-up and Top Down Approaches for Modelling Human Moral Faculties pp. 570-579.
10. McLaren B (2006) Computational Models of Ethical Reasoning pp. 30-32.
11. Guarini M (2006) Particularism and the Classification of Moral Cases pp. 23-26.
12. Muehlhauser L, Louis H (2012) Intelligence Explosion and Machine Ethics.
13. Armstrong S (2015) Consequentialism, Stanford Encyclopedia of Philosophy.
14. Anderson M, Anderson SL (2011) Machine Ethics. Creating an Ethical Intelligent Agent. Ai Magazine 28(4): 15-26.
15. Hibbard B (2011) Model-based Utility Functions.
16. Jean-François B, Iyad R, Azim S (2015) Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars?.
17. Yuchi T, Kexin P, Suman J, Baishakhi R (2018) DeepTest: Automated Testing of Deep-Neural-Network-driven Autonomous Cars.
18. Barrie K (2016) Business opportunities in automated vehicles Journal of Unmanned Vehicle Systems 4(1): 4-6.
19. Everitt T, Filan D, Daswani M, Hutter M (2016) Self-Modification of Policy and Utility Function in Rational Agents.
20. Carrillo H, Dames P, Kumar V, José A. Castellanos (2017) Autonomous robotic exploration using a utility function based on Rényi's general theory of entropy. pp. 42: 235-256.
21. Keren S (2017) Redesigning Stochastic Environments for Maximized Utility.
22. Arindam B, Christopher H, Will N B, Marcus F (2018) Utility function generated saccade strategies for robot active vision: a probabilistic approach. Autonomous Robots 43: 947-966.
23. Bill H (2015) Ethical Artificial Intelligence.

DOI: 10.32474/ARME.2019.02.000136