TOWARDS IMPROVING THE DECISION-MAKING PROCESS OF ARTIFICIAL INTELLIGENCE DEVICES IN SITUATIONS OF MORAL DILEMMAS

Damian Węgrzyn

Polish-Japanese Academy of Information Technology (Poland)

damian@wegrzyn.info

ABSTRACT

Systems using Artificial Intelligence (AI) are created by humans to achieve specific goals. As autonomous decision-making increases, one of the most important considerations is the need to rethink its responsibility. AI's decisions can affect many key aspects of human life. The main issue that arises in the discussion about the usage of AI is the ethical nature of decisions made by AI. Moreover, in some situations these decisions are related to moral dilemmas. This paper deals with problems of moral dilemmas and analyses the moral status of AI devices. As a suggestion to improve the decision-making process in situations of moral dilemmas, the author proposes to apply the fuzzy logic theory. This solution is already used in making choices by AI systems but so far it is not applied to ethical choices in crucial situations of moral dilemmas.

INTRODUCTION

More than 60 years ago, the term Artificial Intelligence (AI) was defined as the tasks performed by a device previously claimed to require human intelligence (McCarthy et al., 1959). Nowadays, the authors attribute a wider area of designation to this concept. The definitions touch upon new skills, such as the ability to flexibly adapt, to learn or even to make decisions based on newly acquired data (Kaplan & Haenlein, 2019). In addition, AI is given the task of adapting through a learning process leading to the ability to sense, reason and act in the most efficient way possible (Tørresen, 2020). After all, it is described as a vague concept with many open questions remaining.

AI is already involved in our everyday life. It has an impact on such important issues as safety, human life and health (Stone et al., 2016). AI can be found in areas such as biomedicine, education, finance, energy, law, space exploration, etc. Good examples of modern applications of AI are aircraft autopilots, where in critical situations the pilot can take control of the machine using his or her experience and acquired skills. In many cases, devices with AI even replace people in making various decisions, such as driving cars, making credit decisions or interpreting medical research results. The rapid development of devices with AI and their entry into everyday life requires them to make decisions. AI systems that make decisions in various areas can affect many key aspects of human life. This raises many questions related to the ethics of decision-making by AI devices. Researchers want to ensure that these systems are ethical, but this is not easy to achieve. Still, the system developers should enable the AI systems to make ethical decisions (Dennis et al., 2015).

Understanding the reasons behind the choices made by modern AI machines is either difficult or sometimes even impossible. This is due to the complexity of the processes that constitute the final choice, such as deep learning using Artificial Neural Networks (ANN). Therefore, there is a need to urgently look at particularly crucial decisions. Teaching the machines morality is undoubtedly a difficult task. Some scenarios cannot be predicted or programmed. Furthermore, there are situations - the so-called moral dilemmas - in which even a man has doubts about what to choose. AI software architecture uses measurable metrics that are not designed for objective moral evaluation. This is because morality is a concept that includes aspects that are not measurable. While the distinction between what is good and evil has at its base an arbitrary or customary set of norms, the definition of many acts is already burdened with subjectivism. In situations of moral dilemma, making choices is often determined by feelings, benefits, or internal prejudices. This is undoubtedly not the case in AI systems. It is known that a machine can only be taught to understand concepts properly if the engineers who design it have a precise definition of the concept. In many cases, the solution is the optimization of decisions, although in real situations it does not always work, because it leads towards the principle of equality and not necessarily justice. In some situations, the usage of rigid algorithms has resulted in discrimination, prejudice or inappropriate choices (Bartneck et al., 2018). As of today, we are unable to teach AI to make fair choices because we do not have an unchanging evaluation within this basic concept of ethics that is out of context or person. In extreme cases, the use of the choices optimization process, without constant analysis and relation to the reality of those choices, may lead to making biased or wrong choices.

Making ethical decisions is a controversial issue, too. When we consider extreme moral dilemmas, in which even people have doubts about the final decisions, we are faced with a problem that is impossible to solve algorithmically, i.e., choosing the lesser evil or the greater good. Values assumed to be immeasurable are elusive for modern technological solutions.

PROBLEMS OF MORAL DILEMMAS

The fundamental problem with ethical dilemmas is whether they exist. Opinions are divided on this point (Holbo, 2002). This paper tends to argue for the existence of moral dilemmas and assumes that the solutions available to the subject are of equal value: none prevails over the other. Moreover, the author supports the thesis that there are no ideal moral theories that would allow one to make ethical choices in every situation. This paper treats a moral dilemma in the strict philosophical meaning. So determined moral dilemma fulfils all the following conditions:

1) an agent is faced with a choice situation between at least two solutions to the problem,

2) each of the solutions to the problem can be chosen by the agent,

3) all solutions are not identical and contradict each other,

4) none of the solutions is subordinate or superior to the other,

5) the agent should select; if there is no choice, one of the available solutions,

6) the agent may choose *n-1* solutions among *n* available solutions,

7) any choice among the available solutions or no choice made brings with it immoral consequences.

Types of moral dilemmas are related to the assumptions made. In general, a distinction can be made between solvable and unsolvable dilemmas. A deeper division involves epistemic dilemmas (one of the solutions takes precedence in a situation) and ontological dilemmas (there are no superior solutions) (McConnell, 2018). This article deals with unsolvable, ontological and symmetrical dilemmas, in the case of which the same moral precept gives rise to conflicting obligations (Sinnott-Armstrong, 1988).

The analysis of the behaviour of AI devices in unforeseen situations introduces uncertainty and imprecision in decision-making. When it comes to situations of moral dilemmas, some of them can be predicted and general scenarios of behaviour or decisions can be prepared for them. Unfortunately, there is a wide spectrum of unforeseeable events in which AI systems will have to make a choice. Then, when deciding or judging, they must be based on the ingrained basic principles of ethics and the data collected so far. It is not known if this ultimately suffices to make choices that can be considered moral.

Defining moral values is a challenge that mankind has grappled with throughout its history. Policymakers and engineers should have methods which allow implementing ethical standards that bring them closer to quantifying ethical values. Finally, let us remember that AI devices are made by people who are subjective and biased in their judgments. By creating ways of ethical choices in situations of dilemma, it is possible to reproduce human faults in AI systems, because ultimately it is a human being who creates AI systems' behaviour and decisions. In this context, the value and cost of both subjective and objective decision-making must be considered. It could be argued that both are needed, albeit to a different extent, to ensure control and balance. Since AI cannot adequately deal with subjectivism, it cannot be deemed to be ethical. However, the actions that will be taken by the AI system are subject to an ethical evaluation. Furthermore, subjectivism is a problem that results directly from human nature and is an inherent factor of choices. The right way to reduce subjectivism and at the same time increase objectivity is to expand the set of choices made in similar situations by people. Crowdsourcing is currently used for this purpose, in which it is assumed that individuals have good intentions and make moral choices. While in known situations it can be inferred with the use of this method in a highly objective manner, the problem remains in new, dilemmatic choices.

Another known problem in evaluating choices is the situational, activity, personal and intentional context. Depending on the context, people evaluate specific facts and make decisions. In the case of an intentional context, morality imposes the choice of good or less evil, in accordance with the current state of knowledge of the decision-maker. The situational and activity context indirectly affects the decisions made, as they may have the so-called mitigating effects. In the case of the personal context, the problem becomes multidimensional. On the one hand, current standards and norms do not allow AI systems to make significant choices based on personality attributes (Di Fabio et al., 2017). On the other hand, there may be situations where the lesser evil is chosen based on the personal context, such as in the trolley problem (Thomson, 1985). Moreover, the context often depends on other phenomena or may even constitute a tangle of events. In such a situation, it is almost impossible to predict the situation or program the scenario.

An important problem with morality in general is the imprecision in defining and assessing moral attitudes. This is because in the final assessment, the problems mentioned so far, such as subjectivism, context or a combination of events, participate to a different extent. In such complex situations people find it difficult to make an honest judgment. The variety of assessments may result, inter alia, from the fact that each of the factors constituting a judgment or choice receives a different weight.

MORAL STATUS OF ARTIFICIAL INTELLIGENCE DEVICES

Nowadays, developers provide AI devices with specific decision rules in situations of moral dilemmas. This requires establishing and defining ethical norms of behaviour in specific difficult situations. The mere implementation of current available and determined indicators allowing to define ethical values is not sufficient for AI systems to decide ethically in all situations. To train ANN models, a large set of unambiguous examples should be collected. For the model to be properly trained and the output to be predictable, as many human judgments as possible must be gathered. By design, this involves showing AI devices clear answers and decision rules to the potential ethical dilemmas they may encounter. It comes down to establishing the most ethical course of action in a difficult situation. Only in this way is it possible to increase the objectivity of decisions due to the variety of situations.

Crowdsourcing is used for such purposes, especially when designing autonomous AI devices, assuming no one is deliberately suggesting the wrong solution. However, in unpredictable new cases AI systems are on their own and have to make choices.

There are exceptions to the rule, which represent the deliberate unethical (in the usual sense) usage of AI devices. One of them is the practice of using IT systems by modern developers for such purposes as lethal autonomous weapons systems (LAWS), drones - killers. The development of information technology proves that war is an important engine of technological progress. It is for military needs that new projects and technologies are constantly being developed, which are ethically controversial.

In terms of the moral status of AI, it is widely assumed that modern AI systems do not have moral status - they are amoral (Bostrom & Yudkowsky, 2014). To categorize a being as having a moral status requires it to belong to a kind that has a sense of sensitivity or reason in the normal way. This can only be done concerning an entity for whom there is no doubt of having an independent moral status (Warren, 1997). If AI devices as self-learning and self-modifying beings will have a sense like the human mind, special attention should be paid to its initial state, as this may have permanent effects and negatively affect its further, ethical self-development, and thus cognitive functions of good and evil (Omohundro, 2008). It is known from the history of ethics that in different periods of mankind the concept of ethics and basic ethical norms have evolved. Once upon a time, slavery was the norm. Quite recently, in the 19th century, women were generally denied the right to vote, as were people of another colour of skin. These days this is categorized as discrimination or even racism (Bostrom & Yudkowsky, 2014). There is a chance that with such a dynamic technological development, AI systems will acquire a moral status, and even become an interpretation of ethics - as an entity with the ability to objectively judge, better than one person, e.g. a judge issuing a judgment in a case.

There are some complications in the direction of AI algorithms towards human thinking. They can fulfil certain social roles, which implies new design requirements such as transparency and predictability. Sufficiently broadly targeted AI machines can operate in unpredictable contexts, requiring security and engineering to incorporate ethical aspects.

Fundamental current issues connected with AI device ethics are transparency, privacy, and awareness of AI (Green, 2017). The decision-making process of AI systems with the complex structure of ANNs is not transparent. Therefore, it is not known on what basis the machine made a specific decision and is not able to explain it. This is known as the black box problem (Winfield, 2017). Nonetheless, AI system designers should make AI device decisions more transparent in an ethical context. Full transparency cannot be ensured, but there is room for greater transparency on how to get closer to quantifying ethical values in programming and determine the choices ultimately made by AI.

It is unacceptable to justify AI's incorrect behaviour by doing nothing about it. By detailing the decision possibilities that AI can make, it allows us to avoid uncontrolled and dangerous decisions of AI systems, especially in situations of ethical dilemmas. Creating algorithms that define a set of ethical values, which are the premises for making AI decisions, will also serve to avoid significant harm. Since people learn moral principles, it must be assumed that systems with AI can also follow unethical paths unconsciously and unintentionally. This requires engineers to constantly improve their moral definition and try to quantify it, which is extremely difficult. In the history of ethics, researchers have attempted to quantify non-measurable attributes to determine the moral status of a given act. A notable example is Bentham's ethical account (Brunius, 1958). This theoretical algorithm of human action describes it as the pursuit of pleasure and avoidance of pain - in line with the hedonistic postulates. The calculations were based on a vector of seven variables (intensity, duration, certainty, speed of occurrence, efficiency, purity and scope). Bentham also defined many kinds of pleasure and distress that a person chooses in certain situations. To evaluate the moral act, the function of the amount of pleasure or pain induced was used. The proposed measurement method significantly simplified the concept of human nature: it reduced the multidimensional complexity of human action to a binary system in known situations. Such a simplified categorization of two values included the hierarchy of not only ethical values but also aesthetic, cognitive and material ones.

AI devices are incapable of moral behaviour. It is their creators who must lay the foundations for their understanding of morality - how to analyse it discreetly. The history of ethics shows that it is not easy to define and quantify such concepts. Ultimately, it is impossible to implement the entire morality in the behaviour of AI devices in a situation where there are no unambiguous and measurable attributes of this issue. Nevertheless, apart from the implementation of ethical behaviour, we leave AI the right to decide also in critical situations.

DISCUSSION ABOUT STANDARDS, NORMS AND RESPONSIBILITIES

To integrate moral or social values with the technological development of AI at all its stages: design, analysis, construction, implementation and evaluation, well-thought-out standards, methods and algorithms are necessary. The idea behind these recommendations is that such devices should be able to make ethical decisions based on a general ethical framework. There is a growing desire among AI engineers for these technologies to be fair and ethical. Standards and norms are usually developed by experts in many areas, which guarantees that it will be an ethically acceptable process.

Recommendations for AI developers

The area of computer ethics' interest is the ethical guidelines for machines with AI. Since ethics should constitute the basis of standards, it is impossible to omit them in this discussion. The list of the current official recommendations in the literature that deal with AI's ethical dilemmas includes the standards of the European Union's Roboethics Special Interest Group (Veruggio, 2006), South Korean Robot Ethics Charter (Korea's Ministry, 2012), reports of German Ethics Commission (Di Fabio et al., 2017), the BS8611 standard by British Standards Institute (British Standards Institute, 2016), and IEEE's Ethically Aligned Design (IEEE, 2019).

The first three basic principles of ethics in the process of creating machines were introduced by Isaac Asimov in the 1940s, known as Asimov's Laws (Asimov, 1942). The first law is principal and covers the protection of human health and life through devices. The other two laws are only a supplement to the first, as they regulate the behaviour of robots concerning the implementation of human commands and the validity of the device's survival. Additionally, the third law tracks the decisions AI devices make - a machine cannot put its existence ahead of human health.

The European Union has secured the ethics of AI machine development with the establishment of the Roboethics Special Interest Group. Article number 1.1 of E.U. Standards (Veruggio, 2006) is concerned with the safety and autonomy of robots. It recommends that the robot have its operators who should be able to limit the autonomy of devices in situations, where their behaviour cannot be guaranteed. This also includes decisions made in situations of ethical dilemmas. This standard should be implemented in all types of robots.

A good example of a standard that describes recommendations for decisions made by AI devices is the South Korean Robot Ethics Charter (Korea's Ministry, 2012), which was established in 2006 and updated in 2012. In the part describing manufacturing standards, it is recommended that in critical situations, AI devices should be prepared for human control. Robot manufacturers should be mindful of minimizing the risk of death or injury to the user, as well as keeping the community safe. In addition, the topic of antisocial and sociopathic behaviour by robots is discussed to minimize the risk of psychological injury to humans. In the second part, dealing with the rights and obligations of users, the standard guarantees them the right to use the robot without risk or fear of physical or mental harm and the right to take control of the robot. The Charter also gives users the right to use the robot in any way if it is fair and within the law. Separately, it is mentioned that the user is not allowed to use the robot in a way that causes physical or mental harm. In the third part, concerning the rights and obligations of the robot, a clause has been added that the AI device cannot deceive a human, and therefore its decision-making should be clear, obvious and transparent in this aspect.

The German Ethics Commission took up the issue of moral dilemmas in the decisions of AI devices. The guidelines were published in 2017 in the report for Automated and Connected Driving (Di Fabio et al., 2017). Clause 5 recommends that AI devices should be designed to avoid critical situations. The authors consider moral dilemmas as situations, in which a machine must choose between two unethical outputs, between which there is no compromise. Thus, it is proposed to continuously develop the entire spectrum of technological options that will allow for anticipation and decision-making with the least possible risk to humans, thereby increasing safety. If, on the other hand, there is a critical situation that cannot be avoided by using available technological solutions, first of all human life should be protected. Therefore, the AI systems must be programmed to accept damage caused to animals or property in a dilemma situation, leading to the risk of human health and life. Extreme dilemma decisions, in which there is a choice between one human life and another, depend on the specific situation and cannot be uniquely standardized or programmed. There is no standardization of the effect’s assessment of decisions, which would be equivalent to a person's moral capacity to make judgments under certain circumstances and historical data. The publication emphasizes that transforming such processes into abstract or general ex-ante evaluations in the form of appropriate programming activities is extremely difficult. For this reason, it is recommended that independent institutions systematically process the lessons learned from the behaviour of AI devices. Decision-making in moral dilemmas cannot be based on personal characteristics such as gender or age. AI programming should be based on the principle of reducing the number of injuries. AI systems also cannot make decisions related to the sacrifice of the other party. Clause 16 differentiates between a fully autonomous system and a system that can be nullified. The second type should be designed in a way that allows for an unambiguous assignment of responsibility: whether it is on the side of the AI system or the side of the user. Clause 18 allows self-learning systems to be implemented only when the security requirements are met and without questioning fundamental ethical principles. Connectivity to the scenario databases is also acceptable if there is a security benefit. However, it is recommended to develop an appropriate standard, including acceptance tests, based on a catalogue of scenarios. Ultimately, in critical situations, the AI system must be able to enter the so-called safe condition, without external human assistance.

British Standards Institute published in 2016 the BS 8611 standard (British Standards Institute, 2016) containing guidelines for the safe design and use of robots. This standard guides to help eliminate or reduce the risk of ethical risks associated with the use of robots. The analysis was based on the standards related to the risk assessment of machines, as well as risk reduction and management. The standard defines various terms, especially ethical harm, ethical threat and ethical risk, which allow for a general understanding of the key and basic principles that determine human behaviour affecting programmed AI devices. A similar approach is presented in the IEEE document published in 2019 (IEEE, 2019). It introduces the vocabulary and models of risk assessment to explain ethical dilemmas.

The responsible innovation

The literature on the ethics of AI devices emphasizes the analysis of responsibility (Dignum, 2017). The authors' conclusions boil down to the issue called responsible innovation (Wong, 2016). This idea assumes that the responsibility for AI machines also rests with manufacturers. This is consistent with the thesis that AI's responsibility is fundamental. P.H. Wong (Wong, 2016) notes that creating AI should be more like raising a child than programming an application.

Responsibility for the development of AI devices is to ensure compliance with basic human principles and values to ensure order and prosperity in a sustainable world. In this context, the creation of AI machines, as an element of responsible innovation, consists of ethics by, in, and for design. V. Dignum (Dignum, 2018) defines ethics by design as integrating the ability to ethically reason in an algorithmic way into the behaviour of AI systems. Ethics in design includes regulatory and technical methods to support the evaluation of the ethical consequences of AI devices that participate in human social structures. Ethics for design assumes a close relationship between developers and users at all life cycles of AI systems in the form of codes of conduct, standards, or certification processes. P. Vamplew et al. (Vamplew et al., 2018) raise issues of legal, ethical and security frameworks that are not sufficient for multi-purpose decision-making by AI systems. The authors propose the paradigm of the multiobjective maximum expected utility. It is based on a combination of vector tools and a non-linear selection of activities, allowing to determine the current effectiveness of the maximum expected utility. T. Arnold and M. Scheutz (Arnold & Scheutz, 2018) propose a scenario generation mechanism that allows to verify the decisions of AI systems in the virtual world to avoid them in the real world. V. Bonnemains et al. (Bonnemains et al., 2018) analyse the ethical reasoning of AI systems. The authors propose an automatic process of judgment of decisions from an ethical point of view, based on models of ethical principles and formal tools describing the situation. To answer a specific ethical dilemma and its moral assessment, the authors use modelling in three ethical areas: utilitarian ethics, deontological ethics and the doctrine of double effect.

Currently, there are many projects in the field of AI ethics development, having a significant impact on the analysis of moral dilemmas of AI devices. They support cooperation in the field of AI ethics (Partnership on AI project) and deal with ethics in autonomous systems (IEEE Ethics in Action in Autonomous and Intelligent Systems project, IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems project). Some projects directly contribute to improving the quality of ethical decisions through crowdsourcing (Moral Machine project).

PROPOSAL OF THE FUZZY LOGIC APPLICATION

On the one hand, ethical standards and recommendations suggest that AI systems should not be guided by individual characteristics, such as age, gender, or physical or mental constitution in critical choices between the preservation of two persons' lives (Di Fabio et al., 2017). On the other hand, there should be a choice of the lesser evil or the greater good. One of the potential possibilities is a hierarchical order of solutions or values. Then the choice seems obvious but it is not reliable. In the case of moral dilemmas, although the available solutions are contradictory, they are nevertheless on the same moral level. The hierarchy of solutions leads to a situation where one of the choices will be morally inferior, which excludes the attributes of a moral dilemma.

In decision-making processes, the ability to deal with uncertainty and imprecision is an important issue and affects the quality of decisions made. Imprecise concepts are attributes of human judgment that are reflected in the process of AI systems programming. Therefore, a mathematical formulation with precise values cannot describe and predict a realistic decision-making process (Xue et al., 2017). To describe fuzzy attributes, a fuzzy inference system can be used, mapping the relationships between many decision components. In such a process of inference, the use of the fuzzy logic theory may be helpful.

The fuzzy logic between the two extreme states 0 and 1 assumes many intermediate values that determine the extent to which the element belongs to the fuzzy set (Zadeh, 1965). In the discretization of fuzzy concepts, all states should be assigned discrete values. Fuzzy logic is currently used in control systems, evolutionary algorithms and neural networks, based on which it is possible to build decision-making systems that analyse ambiguous or sometimes even contradictory features. Nowadays, the fuzzy logic theory is at the base of the decision-making process. There are examples of AI systems based on fuzzy logic in the literature. For instance, the proposal of a pedestrian recognition model that incorporates fuzzy logic into a multi-agent system, to deal with cognitive behaviours that introduce uncertainty and imprecision in decision-making, confirms the high effectiveness of this method (Anderson and Anderson, 2018; Xue et al., 2017). However, these are not ethical decisions. First of all, these methods are based on various personality models that represent features of human nature. In the case of the decision-making process of AI devices, e.g. autonomous vehicles, this is not allowed due to the requirements of applicable standards. In the case of ethical decisions, sets defining unmeasurable values (e.g., from crowdsourcing) can be used to help describe imprecise ethical concepts, such as evil, good, justice, or freedom but it cannot be the final criterion, due to the subjectivism of individual human assessments, mistakes in the answers given or even the immoral goals of individuals. Still, the main advantage of such a solution is the possibility of modelling ethical behaviour simulating human nature, which is ambiguous and imprecise.

The fuzzy set theory in the decision-making process of AI systems in situations of ethical dilemmas can be used in conjunction with many available decision support methods (Ogryczak, 1997). One possibility is the concept of decision preferences. The concept of the preference relationship is currently the basis for researching decisions of individuals. Direct relationship measurement is a difficult task. Preferences are characterized as a binary relation referring to vectors describing multidimensional objects. In formal terms, preferences are usually a preorder or a linear order, i.e. a reflexive, transitive, and consistent binary relation. The relation of preferences enables the decision-maker to be assigned an individual scale of preferences, on which profiles can be evaluated and choices optimized. The function of assigning a value to individual preferences is an ordering function that introduces an order (Bąk, 2013). At this point, to apply the concept of preferences in the decision-making process in moral dilemmas, the function of belonging to fuzzy sets can be used, which will allow for a more realistic moral evaluation of the choice. Such fuzzy inference also considers the features of the decision-maker and can generate different decision preferences, because fuzzy relationships between decision preferences are determined by the fuzzy inference system. Inference in situations of moral dilemmas is a multi-criteria inference, where different solutions may have different vectors of moral evaluations. In addition, there is no general or a priori formulated function. Therefore, the incomparability of individual solutions in the sense of the model does not mean that they are incomparable or indistinguishable. Sometimes it is assumed that in such a situation the set of solutions to the problem is the whole set of effective solutions (Ogryczak, 1997).

The suggestion of using fuzzy logic supports both the objective and subjective approach, because on the one hand it is based on previously known standards, norms or crowdsourcing, and on the other hand, it analyses the decision-making preferences of the decision-maker concerning the dynamic situation, effects and context. Thus, currently used mathematical decision-making mechanisms can be used in situations of unsolvable and ontological dilemmas.

CONCLUSIONS

Nowadays, IT systems make decisions in various areas of everyday life, or they will do so soon. The presented analysis of the complexity of the choices made by AI devices shows the need to increase human safety and brings AI judgments closer to objectivity and ethical behaviour. Teaching the AI devices morality is undoubtedly a difficult task because of its immeasurable character. This paper sets out a possible direction that integrates fuzzy logic theory into the decision-making process of AI systems where there is uncertainty and imprecision.

This paper contributes to the ongoing debate on the automation of ethical decision-making through AI. It shows the importance of this issue and outlines the direction of further research in the moral dilemmas area. The author indicates the importance and complexity of the problem. The presented deliberation of issues related to moral dilemmas in the area of AI is an incentive for further analysis, research and implementation.

ACKNOWLEDGMENTS

The author would like to show his sincere appreciation for helpful comments and kind suggestions for the manuscript and a great deal of encouragement from Professor Alicja Wieczorkowska of the Polish-Japanese Academy of Information Technology, Poland.

This paper was partially supported by research funds sponsored by the Ministry of Science and Higher Education in Poland.

KEYWORDS: moral dilemmas, AI devices, decision-making process, fuzzy logic.

REFERENCES

Anderson, M. & Anderson, S.L. (2018). GenEth: A general ethical dilemma analyzer. *Paladyn, Journal of Behavioral Robotics*, 9(1), 337-357. https://doi.org/10.1515/pjbr-2018-0024

Arnold, T. & Scheutz, M. (2018). The “big red button” is too late: an alternative model for the ethical evaluation of AI systems. *Ethics and Information Technology*, 20, 59-69. https://doi.org/10.1007/s10676-018-9447-7

Asimov, I. (1942). Runaround. *Astounding Science Fiction*, 29(1). Retrieved from http://www.isfdb.org/cgi-bin/pl.cgi?57563

Bartneck, Ch., Yogeeswaran, K., Ser, Q.M., Woodward, G., Sparrow, R., Wang, S. & Eyssel, F. (2018). Robots and Racism. *Proceedings of 2018 ACM/IEEE International Conference on Human Robot Interaction (HRI’18)*, 1-9. Retrieved from https://doi.org/10.1145/3171221.3171260

Bąk, A. (2013). *Microeconometric methods of researching consumer preferences using the R program.* Warsaw: C.H. Beck publishing.

Bonnemains, V., Saurel, C. & Tessier, C. (2018). Embedded ethics: some technical and ethical challenges. *Ethics and Information Technology,* 20, 41-58. https://doi.org/10.1007/s10676-018-9444-x

Bostrom, N. & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish & W. Ramsey (Eds.), *The Cambridge Handbook of Artificial Intelligence* (pp. 316-334). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139046855.020

British Standards Institute (2016). *BS 8611:2016. Ethical design and application of robots*.

Brunius, T. (1958). Jeremy Bentham's Moral Calculus. *Acta Sociologica* 3(1), 73-85. https://doi.org/10.1177/000169935800300107

Dennis, L.A., Fisher, M. & Winfield, A.F.T. (2015). *Towards verifiably ethical robot behaviour*. Retrieved from http://arxiv.org/abs/1504.03592

Di Fabio, U., Broy, M. & Brüngger, R.J. (2017, June). Ethics commission automated and connected driving. *Federal Ministry of Transport and Digital Infrastructure of the Federal Republic of Germany*. Retrieved from https://www.bmvi.de/SharedDocs/EN/publications/report-ethics-commission-automated-and-connected-driving.pdf

Dignum, V. (2017). Responsible autonomy. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI'2017)*, 4698-4704. https://doi.org/10.24963/ijcai.2017/655

Dignum, V. (2018). Ethics in artificial intelligence: introduction to the special issue. *Ethics and Information Technology,* 20, 1-3. https://doi.org/10.1007/s10676-018-9450-z

Green, B.P. (2017, November, 3-4). Some Ethical and Theological Reflections on Artificial Intelligence. In *Graduate Theological Union, Pacific Coast Theological Society*, Berkeley. http://doi.org/10.12775/SetF.2018.015

Holbo, J. (2002). Moral Dilemmas and Deontic Logic. *American Philosophical Quarterly,* 39, 259-274.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019). *Ethically aligned design*. Retrieved from https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf

Kaplan, A. & Haenlein, M. (2019). Siri, Siri in my Hand, who's the Fairest in the Land? On the Interpretations, Illustrations and Implications of Artificial Intelligence. *Business Horizons,* 62(1), 15-25. https://doi.org/10.1016/j.bushor.2018.08.004

Korea's Ministry of Commerce, Industry and Energy (2012). *South Korean Robot Ethics Charter*. Retrieved from https://akikok012um1.wordpress.com/south-korean-robot-ethics-charter-2012

McCarthy, J.J., Minsky, M.L. & Rochester, N. (1959). Artificial intelligence. *Research Laboratory of Electronics Progress Report No 53.* http://hdl.handle.net/1721.1/52263

McConnell, T. (2018). Moral dilemmas. In: E. N. Zalta (Eds.), *The Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/archives/fall2018/entries/moral-dilemmas

Ogryczak, W. (1997). *Multi-criteria linear and discrete optimization: models of preferences and applications to support decisions.* Warsaw: University of Warsaw Press.

Omohundro, S.M. (2008). The Basic AI Drives. In *Proceedings of the 2008 conference on Artificial General Intelligence*. IOS Press, 483-492. https://doi.org/10.5555/1566174.1566226

Project of IEEE Ethics in Action in Autonomous and Intelligent Systems: https://ethicsinaction.ieee.org

Project of IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems: https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

Project of Moral Machine: https://www.moralmachine.net

Project of Partnership on AI: https://www.partnershiponai.org

Sinnott-Armstrong, W. (1988). *Moral Dilemmas.* Oxford: Basil Blackwell.

Stone, P., Brooks, R., Brynjolfsson, E., Calo, R., Etzioni, O., Hager, G., Hirschberg, J., Kalyanakrishnan, S., Kamar, E., Kraus, S., Leyton-Brown, K., Parkes, D., Press, W., Saxenian, A., Shah, J., Tambe, M. & Teller, A. (2016). Artificial Intelligence and Life in 2030. *One Hundred Year Study on Artificial Intelligence: Report of the 2015–2016 Study Panel,* 52.

Thomson, J.J. (1985). The Trolley Problem. *Yale Law Journal,* 94, 1395-1415.

Tørresen, J. (2020, January 7). AI Ethics: How to achieve ethically good artificial intelligence research and development. *AI Ethics Seminars at Chalmers.*

Vamplew, P., Dazeley, R., Foale, C., Firmin, S. & Mummery, J. (2018). Human-aligned artificial intelligence is a multiobjective problem. *Ethics and Information Technology,* 20, 27-40. https://doi.org/10.1007/s10676-017-9440-6

Veruggio, G. (2006). The EURON Roboethics Roadmap, *6th IEEE-RAS International Conference on Humanoid Robots*, 612-617. https://doi.org/10.1109/ICHR.2006.321337

Warren, M.A. (1997). *Moral Status: Obligations to Persons and Other Living Things. Issues in Biomedical Ethics*. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198250401.001.0001

Winfield, P.A. (2017). *ELS issues in robotics and steps to consider them. Part 3: Ethics.* Retrieved from https://www.eu-robotics.net

Wong, P.H. (2016). Responsible innovation for decent non liberal people: a dilemma? *Journal of Responsible Innovation*, 3(2), 154-168. https://doi.org/10.1080/23299460.2016.1216709

Xue, Z., Dong, Q., Fan, X., Jin, Q., Jian, H. & Liu, J. (2017). Fuzzy logic-based model that incorporates personality traits for heterogeneous pedestrians. *Symmetry*, 9(10), 239. https://doi.org/10.3390/sym9100239

Zadeh, L.A. (1965). Fuzzy sets. *Information and Control,* 8, 338-353. https://doi.org/10.1016/S0019-9958(65)90241-X