ISSN 2277-2685
IJESR/July. 2018/ Vol-8/Issue-3/1-10
Mr. G. Venkateshwarlu, / International Journal of Engineering & Science Research

Artificial Intelligence Ethics

Mr. G. Venkateshwarlu, MCA, MTech *1, Mr. C. Santhosh Kumar Reddy, MCA *2, Mr. Ananda Reddy. Y, MSc(Maths); MSc(Stats) *3

*1 Faculty, Dept. of Computer Science, Siva Sivani Degree College, Kompally, Secunderabad – 500100
*2 Faculty, Dept. of Computer Science, Siva Sivani Degree College, Kompally, Secunderabad – 500100
*3 HOD, Dept. of Statistics, Siva Sivani Degree College, Kompally, Secunderabad – 500100

Abstract

In many ways, artificial intelligence (AI) is becoming a constant in our lives today, often going unnoticed even by the well-informed. Apart from the apparent advantages of artificial intelligence, there have sadly been documented instances of individuals experiencing issues as a result of biased decisions made by AI programs. As such, it is imperative to inquire about ethical concerns and the reliability of AI thought about. In this way, the article offers an analysis of the moral considerations that AI application developers ought to make. The European Union's proposed regulations for a reliable and moral AI are presented. The development of moral AI applications is covered.

Keywords

Ethics, Artificial Intelligence, Machine Learning, Deep Learning, Neural Networks, Human-AI interaction

1. Introduction

One could argue that we have entered the artificial intelligence (AI) era. AI technology is being used in more and more aspects of our daily lives, whether we are conscious of it or not. Artificial Intelligence (AI) is present in our "smart" phones and TVs; e-commerce sites and other businesses use AI for personalized product recommendations; natural language processing for automatic translation; voice recognition and generation; and intelligent assistants from Google, Amazon, Apple, and Microsoft (e.g., Alexa, Siri, Cortana) are just a few well-known uses of AI computer programs. Furthermore, artificial intelligence (AI) is currently used in a wider range of decision-making processes, such as those that directly impact human lives and approve bank credit or permit

*Corresponding Author

conditional release from jail (Tolan et al., 2019). Self-driving automobiles and sophisticated domestic robots to assist the elderly and disabled are anticipated in the near future. In this context, human-computer interaction now encompasses human-artificial intelligence interaction to a significant extent.

Apart from artificial intelligence, which is more pervasive in our daily lives, the internet has already turned into a daily necessity as a means of communication, information exchange, entertainment, and online shopping. The internet altered people's habits, behaviors, and mindsets all at once. While there are benefits to using online communication instead of in-person interactions (such removing the need for travel and cutting down on distance), there are also repercussions that are hard to predict at this time. For instance, virtual interactions that includes little to no face-to-face eye contact remove guilt and other constraints, leading to a clear shift in accepted behavioral norms. Furthermore, it can be difficult to figure out who the actual dialogue partner is, and it's possible that we are unaware that we are conversing with a machine.

Important ethical questions are raised by the use of AI to make critical decisions about people, the possibility of discussing with conversational agents while we are unaware, the use of autonomous vehicles or robots on public roads, and the placement of people with disabilities in the care of unassisted robots. Deep neural networks and other machine learning (ML) applications deserve particular attention because, despite their current exceptional performance, they lack a crucial human characteristic that sets humans apart: the ability to rationally explain why they choose one course of action over another. These days, the subject of explainable AI (XAI) is a "hot" topic in AI. Furthermore, judgments made by machine learning algorithms may be biased in light of the data's content because these algorithms rely on statistical techniques that start with vast volumes of data. ML-based natural language processing (NLP) technology can be used to create profiles of any individual from writings shared on social media platforms and emails, which can be abused by taking advantage of a person's vulnerabilities, addictions, or private information.

The aforementioned factors all serve as justifications for looking into the moral implications of AI use in human contact, the ways in which AI-provided data may negatively impact human lives, and what can be done to stop these unethical outcomes. Important businesses like Orange

(Cousson-Postoarca, 2019) and IBM (Banavar, 2016; IBM, 2019) take the ethical questions generated by AI extremely seriously. The High-Level Expert Group on Artificial Intelligence (HLEG) of the European Commission has released an Ethics Guide for Trustworthy AI (AI-HLEG, 2019d). Additionally, the European Commission's "White Paper on Artificial Intelligence" (2020a) and "Report on the safety and consequences for liability arising from robotics, IoT, and artificial intelligence (European Commission, 2020b). Similarly, the European Parliament released a paper emphasizing the necessity of an AI strategy that is human-centered (European Parliament, 2019).

AI ethics should be discussed from the standpoints of developers and the policies that ensure it, at the very least. The affirmation of moral behavior and relationships is a relevant and significant truth. The report goes on to include a section that outlines suggested EU regulations to guarantee that AI applications adhere to ethics. The third section addresses the inclusion of ethical considerations in the applications created by AI programmers. A conclusions section concludes the paper.

2. EU documents for an ethical and trustworthy AI

As the introduction stated, the European Commission, the European Parliament, the Council of Europe, UNESCO, and major corporations (Banavar, 2016; IBM, 2019; Cousson-Postoarca, 2019), as well as other organizations (AI-HLEG, 2019d; AI-HLEG, 2020; European Commission, 2019b, 2020a, 2020b; European Parliament, 2019; UNESCO, 2019), view ethics in the context of AI as a topic of great concern. The European AI HLEG expert group four ethical guidelines have been established by the Commission (AI-HLEG, 2019a):

(i) regard for individual autonomy;
(ii) protection from damage;
(iii) equity; and
(iv) explicability

The Artificial Intelligence (ALTAI) document (AI HLEG, 2020) outlines seven conditions that should be taken into consideration for the creation of AI applications, in addition to these ethical standards and most likely also for enforcing them (particularly the second one):

(i) human participation and observation;
(ii) safety and robustness of the technology;
(iii) adherence to data governance and privacy;
(iv) openness;
(v) Responsibility;
(vi) the state of the environment and society;
(vii) equity, diversity, and nondiscrimination.

3. Developing ethical AI

3.1 What is ethics?

Since ethics is inextricably linked to people, we believe that any study of its specifics in the context of AI should begin with a study of how humans conceptualize ethics, how they establish morality, and how they determine what should or shouldn't be done. Throughout history, sociology, religion, and philosophy have all placed a strong emphasis on ethics. There are numerous definitions of ethics that are necessarily impacted by the presumptive ontological viewpoint. The distinction that Annemarie Piper draws between "good" and "well" may be one that is helpful in the context of ethics and artificial intelligence (Piper 1999). In her view, ethics is concerned with the morally upright. She divides ethical theories into two categories: deontological (Kant, Kierkegaard, and Nietzsche), which begin with certain established conceptions or principles, and teleological (Aristotle and utilitarians, for example), which view as standards of behavior evaluation. The following quote describes several viewing ethics, and they may also be found in the methods used to create AI programs that employ a model of
ethics, as Section 3.2 shall demonstrate: "An action is deemed morally good in deontological ethics due to a particular attribute of the action." itself, not because the action's outcome is positive. Conversely, teleological ethics maintains that the value of what an action creates is the fundamental yardstick of morality. Since the fundamental tenet of deontological theories is that actions should comply to rules or laws, they have been referred to as formalistic.

Velasquez and colleagues (1987) described an experiment in which sociologist Raymond Baumhart questioned some businesspeople "What does ethics mean to you?" and included several of their responses to illustrate some ideas that people have about ethics.

3.2 Ethics and AI

The experiment that Velasquez and associates showed shows how widely the responses are when it comes to ethics. They make reference to social elements like laws and behavioral norms as well as subjective elements like feelings and beliefs. Answers 1-3 are largely pertinent to artificial intelligence programs that determine whether a given action meets with ethical standards, in relation to the topic we are discussing. Still, regarding complexity and implementation potential, these three scenarios differ significantly from one another.

The second scenario is arguably the easiest to handle in AI since it requires confirming that AI-generated words or activities comply with legal requirements. Given that AI programs are like bureaucrats, applying rules mechanically (production rules being one of the well-known knowledge representations in AI), the task appears straightforward (Winograd 1987). It could be difficult to codify rules. Moral laws, and justice laws in particular, are subject to many interpretations; context is crucial, and the laws are predicated on ideas that are difficult, if not impossible, to fully define, such as what is morally just or wrong. In actuality, these noted challenges are unique to example 1. Additionally, instance 1 presents an additional issue, Subjectivity: "with what my feelings tell me is right or wrong," as what a community or individual considers ethical may not always be the same goes for other people. This statement directly leads to the third response in the list.

If laws specify the standards of behavior that are recognized by society, which is what ethics is defined as, the issue is simplified to the second instance in the list. However, machine learning could be used to learn, for instance, moral responses in a conversation, if there aren't any clear "written" regulations that govern the acceptable behavior. The significance of comprehending and managing the issue of the morality of responses produced by a conversational agent using machine learning (ML) training was highlighted in the Tay bot3 instance. Microsoft created this bot on

Twitter to engage users in conversation, but it wasn't until 16 hours later that it had to be taken down due to its inappropriate and abusive language, which turned into racist and misogynistic responses.

The following two questions should be addressed by research on AI's ethical aspects: 1) How feasible is it to build robots, agents, or AI programs that take ethical values into account, either explicitly or implicitly? 2) What are the unique ethical considerations when utilizing AI techniques? Only the first point will be discussed in the remainder of this essay.

Several options for introducing ethical considerations into AI have already been put forth. Asimov (1950) presented the three rules of robots in his science fiction novel series, which are arguably the most well-known proposition.

(i) Robots shouldn't hurt people or let a man suffer by doing nothing.
(ii) Robots ought to follow human commands, unless the first law is broken.
(iii) Robots ought to defend themselves, unless the first two laws are broken.

However, these principles can't account for every scenario, as Asimov himself described in his novels (Asimov, 1950, 1958). Sometimes they even result in their violations. Asimov (1958) described a scenario in "The Naked Sun" in which a human kidnaps a robot and uses its arm as a weapon to commit murder. The robot can't follow the first rule, but it abides by the second. Furthermore, AI might not always be able to determine if a particular course of action would be harmful to a person, even when the first law is taken into account.

Some norms, principles or behavioral patterns should either be "built-in" or taught via machine learning for the implementation of Asimov's three laws—or, for that matter, for any AI system that takes ethics into consideration. AI programs' choices, actions, or responses (for conversational agents, inferences) can be either taught or "calculated," which refers to knowledge processing methods unique to the symbolic paradigm of AI.

Anderson and Anderson (2007) divide computer programs using artificial intelligence into those with implicit ethics and those with explicit ethics, based on a viewpoint that is comparable up to a point. They classify as belonging to the first group ethical standards that are "built-in,"

programmed, and integrated by designers. Neural networks and other machine learning systems that are meant to behave morally would also be included here. However, it is crucial to highlight a key distinction: unlike Tay, who was previously covered in this section, it is uncertain if unethical behavior would occur in the context of neural networks or machine learning.

Programs that incorporate explicit ethics convey rules or fundamental principles clearly; they are not "built-in," but rather can be visualized, examined, and enhanced. New principles may also be introduced, and conclusions drawn. One significant benefit is that systems use learned ethical concepts to justify the morality of each given action. In the implicit case, this isn't necessarily the case. Teleological and deontological ethical theories are distinguished, as was mentioned in Section 3, 1 (Piper 1999). Utilitarianism is a theory that infers good and evil from an action's results. Hedonistic utilitarianism, one of its variations, prioritizes pleasure over all other considerations. According to this hypothesis, an AI program is meant to determine the potential impact of an action and the proportion of people who would find its outcome enjoyable (Anderson & Anderson, 2007). It is important to note that this "moral arithmetic" might lead to people making the incorrect sacrifices for the "good" of the majority.

The deontological method is an alternative to the teleological one in that it begins with principles, norms, or laws rather than the outcome of actions. One method for this could be a formalisation, like in deondic logic4, which permits deductions about what is required, prohibited, permitted, and optional.

Anderson and Anderson also discuss an alternative method that separates what we should do from who we should be, using "virtue" as a fundamental term (Anderson and Anderson 2007). They also discuss Inductive Logic Programming, a machine learning technique that may be applied in a specific subject with several prima facie debts, to find ethically important templates in vast amounts of literature (Anderson and Anderson 2007).

4. Conclusion

Technological advancements must always be accompanied by ethical considerations. Researchers and developers in the field of artificial intelligence (AI) technology should look into and take into account in their products the ethical aspects of expanding computer programs with AI, in relation

to virtual communication, social networks, intelligent artificial agents, devices, robots, and autonomous cars. This is similar to the scientists who have contributed to the development of nuclear technologies and have warned about the dangers of the atomic bomb.

References

AI HLEG (2019a) Ethics Guidelines for Trustworthy AI. Available at https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419 (Last accessed: 31 March, 2020)

AI HLEG (2019b) Policy and investment recommendations for trustworthy Artificial Intelligence. Downloaded from https://ec.europa.eu/digital-single-market/en/news/policy-and-investmentrecommendations-trustworthy-artificial-intelligence (Last accessed: 31 March, 2020)

AI HLEG (2019c) A definition of AI: Main capabilities and scientific disciplines. Downloaded from https://ec.europa.eu/digital-single-market/en/news/definition-artificial-intelligencemain-capabilities-and-scientific-disciplines (Last accessed: 31 March, 2020)

AI HLEG (2019d) Ethics guidelines for trustworthy AI. Downloaded from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai, (Last accessed: 25 July 2020)

AI HLEG (2020) Assessment List for Trustworthy Artificial Intelligence (ALTAI) for selfassessment.
Downloaded from https://futurium.ec.europa.eu/en/european-aialliance/ pages/altai-assessment-list-trustworthy-artificial-intelligence (Last accessed: 25 July 2020)

Anderson, Michael, Anderson, Susan Leigh 2007. „Machine Ethics: Creating an Ethical Intelligent Agent", Artificial Intelligence Magazine, 28:4, Winter, 15-26.

Asimov, I., (1950) I, robot, Gnome Press, New York, NY

Asimov, I., (1958) The naked sun, Bantam Books, New York, NY

Banavar, G. (2016) Learning to trust artificial intelligence systems: Accountability, compliance, and ethics in the age of smart machines. Armonk, NY: IBM Research.

Comisia Europeana (2019a) AI - The future of work? Work of the future!. Downloaded from https://ec.europa.eu/epsc/sites/epsc/files/ai-report_online-version.pdf (Last accessed: 31 March, 2020)

Cousson-Postoarca, R. (2019) Ensuring ethical AI is human-centric. Downloaded from https://www.orange-business.com/en/blogs/ensuring-ethical-ai-human-centric (Last accessed: 31 March, 2020)

Deloitte University Press (2017) AI-augmented government Using cognitive technologies to redesign public sector work. Downloaded from https://www2.deloitte.com/content/dam/insights/us/articles/3832_AI-augmentedgovernment/DUP_AI-augmented-government.pdf (Last accessed: 31 March, 2020)

European Commission (2019b) Building Trust in Human-Centric Artificial Intelligence, COM(2020) 168, Available at https://ec.europa.eu/transparency/regdoc/rep/1/2019/EN/COM-2019-168-F1-EN-MAINPART-1.PDF

European Commission (2020a) WHITE PAPER On Artificial Intelligence -A European approach to excellence and trust, COM(2020) 65, Available at https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligencefeb2020_en.pdf (Accessed: 31 March, 2020)

European Commission (2020b) Report on the safety and liability implications of Artificial Intelligence, the Internet of Things and robotics, COM(2020) 64, Available at https://ec.europa.eu/info/sites/info/files/report-safety-liability-artificial-intelligencefeb2020_en_1.pdf (Accessed: 31 March, 2020)Comisia Europeana (2020c) A European strategy for data, COM(2020) 66. Downloaded from https://ec.europa.eu/info/sites/info/files/communication-european-strategy-data-19feb2020_en.pdf (Last accessed: 31 March, 2020)

European Parliament (2019) EU guidelines on ethics in artificial intelligence: Context and implementation. Downloaded from https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/640163/EPRS_BRI(2019)640163_EN.pdf (Last accessed: 31 March, 2020)

IBM (2019) Everyday Ethics for Artificial Intelligence. Downloaded from https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf (Last accessed: 31 March, 2020)

Krzysztof, C. (2018). Deep Neural Networks - A Brief History. Advances in Data Analysis with Computational Intelligence Methods pp.183-200

O'Neil, C (2016) Weapons of Math Destruction, Crown Books

Piper, Annemarie 1999. „Binele". în Schnadelbach, H, Martins, E. (eds.) Filosofie. Curs de baza. Bucure?ti: Editura ?tiin?ifica.

Tolan S., Miron M., Gomez E. and Castillo C. (2019) Why Machine Learning May Lead to Unfairness: Evidence from Risk Assessment for Juvenile Justice in Catalonia.

https://chato.cl/papers/miron_tolan_gomez_castillo_2019_machine_learning_risk_assessment_savry.pdf, (Last accessed: 31 March, 2020)

Trausan-Matu, S. (2003) Psihologia robo?ilor, în G.G. Constandache (ed.), Oglinda con?tiin?ei, Politehnica Press, pag. 186-196.

Trausan-Matu, S. (2013) A Polyphonic Model, Analysis Method and Computer Support Tools for the Analysis of Socially-Built Discourse, Tools for the Analysis of Socially-Built Discourse, Romanian Journal of Information Science and Technology 16(2-3), pp. 144-154

UNESCO (2019) Beijing Consensus on Artificial Intelligence and Education, Downloaded from https://unesdoc.unesco.org/ark:/48223/pf0000368303, (Last accessed: 31 March, 2020)

Velasquez, M., Andre, C., Shanks, T., and Meyer, M.J. (2017) What is Ethics? https://www.scu.edu/ethics/ethics-resources/ethical-decision-making/what-is-ethics/ (Last accessed: 3 March, 2021).