International Journal of User-System Interaction 13(3) 2020, 136-148 

# Ethics in Artificial Intelligence

Ştefan Trăușan-Matu

University Politehnica of Bucharest
313 Splaiul Independenței, Bucharest, Romania

Romanian Academy Research Institute for Artificial Intelligence "Mihai Drăganescu"
Calea 13 Septembrie nr. 13, Bucharest, Romania
*E-mail: stefan.trausan@upb.ro*

**Abstract.** Artificial intelligence (AI) is becoming nowadays a daily presence in our lives, in several ways, many times not noticeable even for informed people. In addition to the evident benefits of AI, there were unfortunately reported cases of people that had problems due to biased decisions provided by AI applications. Therefore, enquires about ethical issues and trustworthiness of AI are very important to be considered. In this sense, the paper presents an analysis of the ethical aspects that should be taken into account by the developers of AI applications. Regulations for an ethical and trustworthy AI, which were proposed by the European Union are introduced. Ways of developing ethical AI applications are discussed.

**Keywords**: ethics; artificial intelligence; machine learning; deep learning; neural networks; human-AI interaction

DOI: 10.37789/ijusi.2020.13.3.2

## 1. Introduction

It may be said that nowadays we entered in the era of artificial intelligence (AI). We use AI technology, with or without being aware about its presence, in an increasing amount of our daily life. Our "smart" phones and TV-s include AI, personalized recommendations for products in e-shops or other businesses use AI techniques, natural language processing for automatic translation, voice recognition and generation, intelligent assistants of Google, Amazon, Apple, and Microsoft (e.g., Alexa, Siri, Cortana) are AI computer programs, to enumerate only several well known applications. Moreover, AI is involved now on a larger scale, including, for example, decisions that directly affect human's lifes, such as approving a bank credit or allowing conditional release from prison (Tolan et al., 2019). In the near future are expected self-driving cars and domestic advanced robots for supporting

elderly people or persons with disabilities. In this context, Human-Computer Interaction includes now in a very important degree *Human-AI Interaction*.

In addition to AI, which is now becoming a constant presence in our lives, internet has already become a daily appliance, as a medium for sharing information, communication, entertainment, and e-commerce. In the same time the internet changed human habits, behaviors and mentalities. Replacing face-to-face with online communication has advantages (such as eliminating the necessity of physical movement, reducing distances) but it also has consequences that are still difficult to anticipate. For example, virtual contacts, with reduced or not existent face-to-face gaze, eliminates restraints, removes or at least reduces shame, driving to evident changes of norms of behavior. Moreover, the real identity of the dialog partner may be hard to detect, including the possibility that we do not realize that we talk with a machine.

The usage of AI for taking important decisions about people, letting persons with disabilities in the care of unassisted robots, the possibility that we discuss with a conversational agent without being aware, and the presence of autonomous cars or robots on the streets rise important ethical questions. A special attention should be paid to the applications that use deep neural networks and other machine learning (ML) applications that now have outstanding performances but lack an important feature of humans that they replace: they cannot explain why they took a decision and not another. This is a "hot" topic in AI now, the so called explainable AI (XAI) problem. Moreover, because machine learning use statistical methods starting from large amounts of data, decisions may be biased depending on the content of these data. From another point of view, the powerful natural language processing (NLP) technology based on ML can be used for constructing profiles of any person from the texts exchanged on social networks and emails, which may be used in illicit ways by taking advantage of one's weaknesses, addictions or personal data.

All the above considerations are reasons for investigating the ethical aspects of human interaction with AI, the way data provided by AI can affect human lifes in incorrect ways, and what should be done for the prevention of unethical effects of using AI. The ethical issues raised by AI are very seriously considered by important companies, such as IBM (Banavar, 2016; IBM, 2019) or Orange (Cousson-Postoarca, 2019). European Commission's High-Level Expert Group on Artificial Intelligence (AI HLEG) has published

an Ethics Guide for Trustworthy AI (AI-HLEG, 2019d), and provided a special attention to ethics in the "White Paper on Artificial Intelligence" (European Commission, 2020a) and in the "Report on the safety and liability implications of artificial intelligence, the Internet of Things and of robotics" (European Commission, 2020b). In the same sense, the European Parliament published a document highlighting the need for a human-centered AI approach (European Parliament, 2019).

The discussion about ethics in AI should be done from at least two perspectives, that of policies for assuring it and that of developers. A related important fact is the validation of ethical acts and interactions. The paper continues with a section that introduces recommended European Union policies for assuring that ethics is respected by AI applications. The third section discusses how ethical issues may be included in the applications by the AI developers. The paper ends with a conclusions section.

## 2. EU documents for an ethical and trustworthy AI

As was mentioned in the introduction, ethics in the context of AI is considered as an important topic of concern, by important companies (Banavar, 2016; IBM, 2019; Cousson-Postoarca, 2019), by the European Commission, the European Parliament, the Council of Europe, UNESCO, etc. (AI-HLEG, 2019d; AI-HLEG, 2020; European Commission, 2019b, 2020a, 2020b; European Parliament, 2019; UNESCO, 2019).

The AI HLEG expert group of the European Commission has identified four ethical principles (AI-HLEG, 2019a):

(i) respect for human autonomy,
(ii) prevention of harm,
(iii) fairness,
(iv) explicability.

In addition to these ethical principles and probably also for enforcing them (and especially the second one), the same group introduced seven requirements that should be taken for the development of AI applications, which have been detailed in the "Assessment List for Trustworthy Artificial Intelligence (ALTAI)" (AI HLEG, 2020):

1. human involvement and surveillance;
2. technical robustness and safety;
3. respect for privacy and data governance;
4. transparency;
5. accountability;
6. the well-being of society and the environment;
7. diversity, non-discrimination, and equity.

The *Human involvement and surveillance* principle states that humans should have the possibility of supervision and control that AI applications do not undermine human autonomy and do not cause certain physical or moral harm (European Commission, 2020a). For example, the passengers of autonomous cars should have the possibility to take control of them in an emergency. Other examples start from bad instances of applying ML, that affected civil rights and drove to the refusal of social security benefits, the rejection of a credit card application (Commission European Union, 2020a) or prison conditional release (Tolan et al, 2019). Such a situation was signalled even in 1980, when an algorithm for admission at a medicine school in London generated a biased decision, incorrecting refusing a variety of students[1]. In these kind of cases, the solutions proposed by AI should only take effect if they are reviewed and validated by a person before application. Machine learning is a powerful technology of AI but it can generate results in a partial or even total autonomous way, a desirable feature in many cases but potentially harmful; the results may be unpredictable and lacking the capacity of explaining the decisions (the XAI problem – see also the transparency and accountability requirements). Therefore, if the results may affect humans in any ways, assessments are needed.

For AI based applications that drive cars, control robots or any other devices, human surveillance is not required, and if desired, it is not always possible or there is a possibility of moments of human inattention. Therefore, extensive tests of their functioning are needed. However, a major problem is the fact that these systems, when they have machine learning capabilities, as mentioned above, they may be autonomous and unpredictable.

*Technical robustness and safety* are requirements of any products and

---

[1] https://spectrum.ieee.org/tech-talk/tech-history/dawn-of-electronics/untold-history-of-ai-the-birth-of-machine-bias, accessed on March 6, 2021

obviously should apply also to AI. There should not be any wrong decisions taken by AI algorithms, faulty actions or accidents in the functioning of the developed AI products, such as robots, autonomous cars, etc. that could harm humans or drive to unethical actions.

*Respect for privacy and data governance* are very important requirements that are directly related to ethics. Especially natural language processing programs, remarkable applications of AI, can be used for illicit purposes, for example, messages exchanged on social networks may provide user data that can be utilized for malevolent or even illegal purposes, for example, bullying, phishing, blackmailing, extorting, etc.

Video surveillance, another AI remarkable achievement, permits remote biometric identification, with obvious advantages for detecting terrorists but also with the possibility of observing and controlling ordinary people, especially in dictatorial countries, but also in any country in illegal purposes. Machine learning can automatically classify people in various purposes but the statistical nature of its processings can, for example, fire employees with outstanding performance but who do not fit on the average, such an example being emphasized by O'Neil (2016) in the field of education in the USA. Other cases of not ethical use of AI in gaining advantages may be found in the commercial field, changing the price of some services (for example, car-sharing, example emphasized by a colleague) taking into account special situations detected by instantaneous data analysis.

*Transparency, accountability* (the ability to be able to give answers about their decisions, to be responsible for the actions or decisions taken), and explainability (XAI) are basic features that should be assured for AI in order to ensure compliance with ethical principles. These are critical issue especially in the case of neural network-based AI technologies, which are similar to a black box that, starting from pairs of input-output data learns but, after the phase of learning, cannot explain why a specific output was generated. Moreover, users should be informed if they are using a device (for example, a vehicle) controlled by a human or an AI, if they are interacting with a person or with a conversation agent having AI. On social networks, for example, it is already a practice that users are sometimes deceived by messages generated by programs with AI. Moreover, these agents may utilize in unethical or even illicit ways the data on users' weaknesses extracted from

their interaction history by natural language processing techniques.

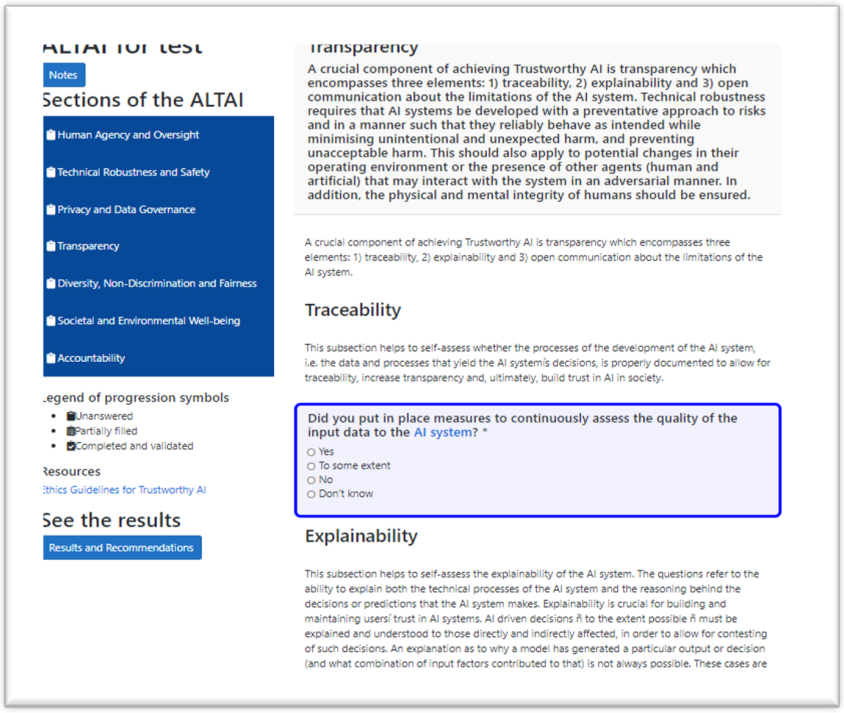


Figure 1 - Verification of traceability related to the transparency criterion of the ALTAI list of verification of ethical principles and trust of AI applications (https://altai.insight-centre.org/AL/119/5, accessed on March 6, 2021)

Anyway, users should be provided with "clear information regarding the capabilities and limitations of AI systems, in particular the purpose for which the systems are designed, the conditions under which they are expected to operate according to the intended purpose and the expected level of accuracy achieve the specified goal" (European Commission, 2020a).

AI HLEG elaborated between June 2018 and June 2020 the ALTAI document (AI HLEG, 2020), which provides lists of questions for each of the seven criteria discussed in the first part of this section. A prototype interactive website was also developed (https://altai.insight-centre.org) where users have to fill questionnaires for answering ALTAI questions in order to check if their developed AI application meets the needed criteria. In Figure 1 it is shown an excerpt from the transparency criterion check.

## 3. Developing ethical AI

### 3.1 What is ethics?

Because ethics is directly related to humans, we consider that any investigation of its particularties in the context of AI should start from an enquiry of how humans conceive it, how they define standards of right and wrong, how they decide what should or not should be done. Along history, ethics was a major subject in philosophy, sociology, and religion. Several definitions of ethics are possible, inevitably influenced by the assumed ontological perspective. A dichotomy that I consider useful in the context of ethics and AI may be that made by Annemarie Piper starting from the analysis of the words "good" and "well" (Piper 1999). Ethics, in her vision, deals with the moral good. She classifies ethical theories into teleological (for example, Aristotle and the adherents of utilitarianism), which consider as criteria of judgement the results of actions and behaviour, and deontological (Kant, Kierkegaard, and Nietzsche), which start from some predefined concepts or rules. The differences between the two types of viewing ethics are described in the following quote and they may be found also in the ways of developing AI programs that use a model of ethics, as will be seen in Section 3.2: "In deontological ethics an action is considered morally good because of some *characteristic of the action itself, not because the product of the action is good*. (…) By contrast, teleological ethics (…) holds that the basic standard of morality is precisely *the value of what an action brings into being*. Deontological theories have been termed formalistic, because their central principle lies in the *conformity of an action to some rule or law*."[2]

For illustrating some visions that humans have about ethics, Velasquez and collaborators (1987) wrote about an experiment in which sociologist Raymond Baumhart asked some bussiness people "What does ethics mean to you?" and included several of their answers:

1. "Ethics has to do with what my feelings tell me is right or wrong."
2. "Being ethical is doing what the law requires."
3. "Ethics consists of the standards of behavior our society accepts."
4. "Ethics has to do with my religious beliefs."

[2] https://www.britannica.com/topic/deontological-ethics, accessed on March, 6, 2021

5. “I don't know what the word means.” (Velasquez et al., 1987)

## 3.2 Ethics and AI

As it can be seen from the experiment presented by Velasquez and collaborators, the answers about ethics differ very much. They refer both to subjective aspects: feelings, beliefs, and to social aspects: norms of behavior and laws. Referring to the topic we are addressing, answers 1-3 are more or less relevant for artificial intelligence programs that decide whether a particular action complies with some rules of ethics. However, there are important differences among these three cases, in terms of complexity and possibilities of implementation.

The simplest case to approach in AI is probably the second, because it needs the verification of the compliance of AI actions or generated text with specified laws. It seems a simple task because AI programs have been compared to a bureaucrat (Winograd 1987), which applies mechanically some rules (and production rules are one of the well known knowledge representation in AI). However, there may be some difficulties because the rules may be hard to formalize. Moral and especially justice laws have many times multiple interpretations, the context is important and they are based on concepts such as what is ethical, good, right, wrong, etc., which are hard to be formalized, being a very difficult task, in general, if not sometimes impossible to complete. These remarked difficulties are in fact specific also to case 1. Moreover, case 1 introduces another problem, subjectivity: “with what my feelings tell me is right or wrong”, because what is considered ethic for a person or for a community might not be the same for others. This remark drives immediately to the third answer in the enumeration.

When ethics is defined as “the norms of behavior accepted by our society”, if the these norms are stated in laws, the problem is reduced to the second case from the list. However, if there are not explicit “written” laws that cover the accepted behavior, machine learning may be used instead for learning, for example, ethical replies in a conversation. The importance of learning and controlling the problem of ethicality of replies generated by a conversational agent trained with ML was emphasized in the case of the Tay bot[3]. Microsoft

[3] https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation, retrieved on March 6, 2021

delivered this bot on Twitter for entering in dialog with people but was forced to remove it only after 16 hours because its language was abusive and offensive, it became racist and mysoginist in its replies.

The goals of investigating ethics aspects of AI should answer to two questions: 1) What are the possibilities of implementing robots, agents or AI programs that consider either implicitly or explicitly ethical principles and how it can be done? 2) What are the peculiarities of ethics in using using AI techniques? In the rest of this paper we will focus only on the first point.

There already have been proposed several possibilities to introduce ethical dimensions into AI. Probably the best known proposal are the three laws of robotics introduced by Isaac Asimov in his series of science fiction novels (Asimov 1950):

(1) Robots should not harm people or, by inaction, to allow a man to suffer.
(2) Robots should obey humans' orders, except when the first law is violated.
(3) Robots should protect themselves, except in cases when the first two laws are violated.

However, as Asimov himself decribed in his novels (Asimov, 1950, 1958), these laws sometimes lead to blockages or even to their violations and cannot cover all possible situations. In "The Naked Sun", Asimov (1958) presented a situation when a robot's arm is taken and used as a weapon by a human for a murder. The robot follows the second rule but cannot obey the first one. Moreover, considering even only the first law, there might be situations when AI cannot infer that a certain action would harm a human.

For the application of the three laws of Asimov and, in fact, for any AI system that considers ethics, some rules, principles or ways of behaviour should either be "built-in" or learned by machine learning. The decisions, actions or answers (in the case of conversational agents) of AI programs may be either as those learned, or "calculated", that means inferred through some knowledge processing techniques specific to the symbolic paradigm of AI.

From a perspective, similar until a point, Anderson and Anderson (2007) classify computer programs with AI into those with implicit ethics and those with explicit ethics. They place in the first category ethical norms that are incorporated by designers, which are programmed, which cannot be modified, which are "built-in". I would add here also neural networks or some

ML systems that are supposed to act ethicaly. Nevertheless, an important difference should be emphasized: in the case of neural networks or ML it is not sure that unethical acts would happen, as was the case of Tay, previously discussed in this section.

In programs with explicit ethics, rules or some basic principles are represented explicitly, they are "built-in" but they can be visualized, analyzed, and improved; inferences can be done and new ones can be added. A major advantage is that systems with explicit ethics principles may explain whether a particular action is good or bad by appealing to memorized ethical principles. This is not always the case in the implicit case,

As was stated in Section 3,1, ethical theories can be divided into teleological and deontological (Piper 1999). A major teleological ethical theory is utilitarianism, in which good and evil are deduced from the consequences of actions. One of its variants, hedonistic utilitarianism, puts pleasure as the most important goal. Anderson and Anderson (2007) say that according to this theory, an AI program it is supposed to calculate what the effect of an action might be and how many people would consider its result as pleasant. We should emphasize here that this "moral arithmetic" can drive to wrong sacrifices of the individuals for the "good" of the majority.

The alternative to the teleological approach is the deontological one, which starts from principles, norms or laws, not from the result of the actions. For this purpose, one approach might be a formalization, for example in deondic logic[4], which allows inferences about what is allowed, forbidden, optional, and mandatory.

Anderson and Anderson mention another approach, which considers "virtue" as a basic concept, deciding what we should do from what we should be (Anderson and Anderson 2007). They mention also Inductive Logic Programming, a ML technique, as a way of searching for ethically relevant templates in large volumes of text, which can be used in a narrow field, characterized by several *prima facie* debts (Anderson and Anderson 2007).

## 4. Conclusions

Consideration of the ethical dimensions must always accompany

[4] https://plato.stanford.edu/entries/logic-deontic/, accessed at March, 6, 2021

technological advances. Similarly with the scientists who have contributed to the development of nuclear technologies and have warned about the dangers of the atomic bomb, researchers and developers in the domain of AI technology should investigate and consider in their products the ethical aspects of expanding computer programs with artificial intelligence, in relation to virtual communication, social networks, intelligent artificial agents, devices, robots, and cars that have autonomy.

As it results from the paper, there are concerns and even official documents in ensuring that robots and AI programs comply with ethical standards. An ideal would be if they could be able to reason about the ethical dimension of the actions taken. However, this is very hard, if not impossible to achieve, in general. A major problem is that not everything can be calculated, that AI has limits. Empathy, awareness, understanding of language, consciousness (and it should be emphasized that consciousness has a major role in humans in judging their own immoral and unethical acts), major characteristics of human life (Trausan-Matu, 2003) are goals still met only partially by AI. All these features are important elements in deciding what is good and what is bad. The specific voices of the humans' inner dialogues involved in empathy and consciousness, from the perspective of the polyphonic model (Trausan-Matu, 2013), can enter into a polyphonic fabric, which potentially may reach the perfection of Johann Sebastian Bach's creations. The harmony of such polyphonies can be seen as an archetype of the concept of good, of the model of co-existence of a human community that is in line with ethical principles. However, it is a question if and how it may be implemented in AI.

## 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-investment-recommendations-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-intelligence-

main-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 self-assessment. Downloaded from https://futurium.ec.europa.eu/en/european-ai-alliance/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 Europeană (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-augmented-government/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-MAIN-PART-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-intelligence-feb2020_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-intelligence-feb2020_en_1.pdf (Accessed: 31 March, 2020)Comisia Europeană (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 bază.* București: Editura Științifică.

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. Downloaded from 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 roboților, în G.G. Constandache (ed.), *Oglinda conștiinței*, 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).