Ana FRICHAND
Biljana BLAZHEVSKA-STOILKOVSKA

UDK: 17.021:159.947.2]:004.8
Original research paper

INTEGRATING MORALITY INTO INTELLIGENT MACHINES – CAN ARTIFICIAL INTELLIGENCE MAKE UNSUPERVISED MORAL DECISIONS?

***Abstract**:*

With the expansion of artificial intelligence and advanced technologies, the world in the 21st century is rapidly changing and imposing new living dynamics. Although such changes affect all age groups, younger generations accept them faster and react more positively. The new cohorts - Generation Z and Alpha - live in a digital world that affect their lifestyle, interpersonal relations, quality of mental health, psychological well-being and everyday challenges. The presence of the so called "Frankenstein effect" in some adults provoked by the fast development of artificial intelligence and robotics, reflects a "humans versus machines" position, viewing artificial intelligence as a threat to humanity. However, the reality is that digital and human world are not in conflict, since many people are already using artificial intelligence tools on daily basis. It is implemented in certain aspects of medicine, education, business, law, agriculture, industry, space technology and in many other fields. With this in mind, the aspect of morality pops up as a very important one. A frequently asked question is: Does artificial intelligence have the capacity to make moral decisions independently? Therefore, integrating morality into AI algorithms is one of the priorities that interdisciplinary teams from engineering and computer sciences, psychology, philosophy, sociology, law, etc. are intensively working on. This paper addresses this issue by presenting findings from relevant research which discusses the challenges and possibilities for integrating the dimensions of morality and introducing human values into autonomous systems that perform complex tasks.

***Keywords**: morality, moral reasoning, moral decision making, artificial intelligence.*

Introduction

Artificial intelligence (AI) today is deeply embedded in various aspects of human lives. With its help, autonomous systems (e.g. autonomous vehicles or robots-assistants) influence development of societies in various ways. They perform complex tasks, while some even interact and collaborate with people. AI's ability to quickly learn and use large databases allows it, for example, to process natural languages expeditiously and efficiently, to analyze data and be efficient in simulations, write financial reports, perform repetitive tasks over a long period of time (important in industrial processes), recognize patterns and details (especially important in medical diagnostics), simultaneously perform multiple tasks (if trained for it) and many other things.

Such expansion and power of AI seems frightening to some people, provoking the so-called "Frankenstein effect" – a fear that human creations will turn against and destroy humanity if science is misused by unscrupulous individuals. At present, one of the most common concerns related to the development of AI, is that its increased use will lead to the closure of many jobs, and thus to the loss of the need for certain professions. For example, Kelly (2024)[1] in the text published this year in "Forbes", writes that certain jobs like those related to data analysis, writing financial reports, routine appointments (such as meetings, medical examinations, etc.), jobs that require memorizing data, basic customer services, as well as repetitive administrative tasks, are among the most affected by the growing trend of using generative AI, because these jobs are highly susceptible to automation. On the other hand, professional roles that require significant social and emotional competencies (for example, psychotherapy, psychological counseling, making complex business decisions, social work, teaching, work in the sales sector where employees enter into direct interactions with clients, as well as managers, lawyers etc.), are less affected due to the crucial role of the human factor in these processes.

Hatzius et al. (2023)[2] in a text published on March 26 last year in *Economic Research* of the Goldman Sachs investment bank, using data on jobs in the United States and Europe, indicate that in 2023 about two-thirds of existing jobs were exposed to some degree of automation with AI and that generative AI can replace about ¼ of them. According to these authors, on a global scale AI could subject to automation about 300 million jobs. In addition, *Resume Builder*[3] analysis from November 8, 2023, report findings from interviews with 750 business

[1] https://www.forbes.com/sites/jackkelly/2024/02/28/what-white-collar-jobs-are-safe-from-ai-and-which-professions-are-most-at-risk/

[2] https://www.key4biz.it/wp-content/uploads/2023/03/Global-Economics-Analyst_-The-Potentially-Large-Effects-of-Artificial-Intelligence-on-Economic-Growth-Briggs_Kodnani.pdf

[3] https://www.resumebuilder.com/1-in-3-companies-will-replace-employees-with-ai-in-2024/

leaders of companies that already use or plan to use AI in 2024. In summary, it shows that 53% of companies in 2023 were already using AI, while 24% declared their plans to start using it in 2024. Moreover, 37% of the companies using AI declared that it replaced some of the jobs in 2023. As for 2024, 44% of companies (both those that are using or are planning to use AI) estimated that some employees will be laid off due to the use of AI.

However, AI is not omnipotent. It has a number of limitations. It is known, for example, that consciousness and conscientiousness are not part of AI. Artificial intelligence also does not possess theory of mind, nor emotional intelligence. When it comes to morality, the latest research show that AI still cannot make unsupervised moral decisions. It also lacks the ability to adapt to completely new situations that require the flexibility of the human mind, for which AI is not trained. Its weakness is the true creativity and innovation that humans possess, which involves thinking outside the box of existing data, etc. Hence, humans despite the rapid development of AI, still remain superior in many aspects, especially in emotional depth (including moral feelings), intuition, the ability for moral reasoning and moral decision-making, creative problem solving and having a rich life experience that allows them to understand and dynamically adapt to the changing context in which they lives.

Artificial intelligence and its various forms

When talking about AI in terms of its ability to learn and apply acquired knowledge, according to the experts at IBM[4], it is commonly referred to three types: weak (or Narrow AI), strong (or General AI - still a theoretical concept) and artificial superintelligence (or Super AI – labeled as "science fiction"). The first type is found, for example, in autonomous vehicles, image recognition software, or virtual assistants. Natural language processing is also a type of AI. The term "weak" indicates that its capabilities are limited to performing specific actions or commands. Large Language Models (LLMs) development enabled a new chapter in the advancement of generative AI and its ability to create various contents similar to those of the real world. Although it has been around as a concept since the 1950s, the development of AI has accelerated since 2012 with the breakthrough of artificial neural networks (the most common type of generative AI), which allow intelligent machines to learn rapidly by simulating how brain processes information[5]. The latest data reported by McKinsey & Company (2024)[6], show that in 2024, generative AI is increasingly applied by organizations (that is, 65% of those included in the survey declared that they regularly

[4] https://www.ibm.com/think/topics/artificial-intelligence-types

[5] https://www.ibm.com/think/topics/artificial-intelligence-types

[6] https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

use it in their work). Thus, as pointed by Bonnefon, Rahwa & Shariff (2024), the progress of AI at present is measured in months, not in years (p.654).

The second type - General or Strong AI - is still a theoretical system in development that could learn unsupervised and perform a wide range of multi-functional activities rivaling the capabilities of humans, which would likely allow it to be the first AI to successfully pass the Turing test[7]. However, experts are divided on the speed at which this type of AI could be achieved. According to Berruti, Nel & Whiteman (2020), some authors believe that there is a possibility that this type of AI will be developed by 2030 or 2040, while others are of the opinion that humanity is not even close to developing it (for example, Rodney Brooks - former professor at MIT and co-founder of iRobot - believes that general AI will not arrive before the year 2300).

Finally, artificial superintelligence is probably what instills the most fear and distrust in people, contributing to the previously mentioned "Frankenstein effect". Namely, this superior type of AI (if ever achieved), will learn so quickly and so independently that it will overcome human cognitive capacities and abilities, becoming many times more powerful and therefore uncontrollable. Such superior AI will be able to autonomously think, reason, make decisions, and so on. It is assumed that it will be a key part of fully self-aware AI and other types of autonomous humanoid robots, becoming the most powerful form of intelligence on the planet[8]. However, all this is still a speculation.

In terms of its functionality, AI comes in four forms[9]: reactive machines, limited memory, theory of mind AI and self-aware AI. Thus, when social media algorithms suggest us what to watch next based on our browsing history (e.g. cooking apps), we are dealing with reactive machines that are AI systems designed to perform specific tasks. Artificial intelligence with limited memory, unlike reactive machines that have no memory, can use past and current data in order to provide guidance on how to reach a desired outcome. Theory of mind AI and even more so self-aware AI are still non-realized forms. However, only theoretically, if an AI possesses a theory of mind, it could understand people's thoughts and feelings, as well as predict actions based on the other entity's known preferences. Hence, it could simulate human-like relationships. Incorporating emotions into theory-of-mind AI is the next step researchers are targeting, trying to enable such AI to understand and respond appropriately to different emotional states in humans by analyzing various relevant data. The most distant and still unattainable form of AI - the self-aware AI – which is a notable characteristic of superintelligence, in addition to people's thoughts and emotions, could also understand its own internal states and traits, and would possess its own thoughts, emotions, needs and beliefs. Yet today, AI still does

[7] https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-artificial-general-intelligence-agi#/

[8] https://www.ibm.com/think/topics/artificial-intelligence-types

[9] Ibid.

not have the capacity to understand or respond adequately to human emotional states and feelings.[10]

Artificial intelligence and social robotics

Artificial intelligence and robotics—as a branch of engineering and computer science where machines are programmed to perform highly specialized tasks—are two completely different disciplines. In reality, there are robots with AI, but also those that do not actually include AI (which are currently more numerous). It is assumed that in the future, as AI becomes increasingly sophisticated, designs where it will be integrated into robots will be more common. Today, the application of AI in robotics is seen in industrial production, various types of businesses, medicine, agriculture, space technology, military industry, etc. (Martin, 2021)[11]. Another domain in which these two disciplines are intertwined is that of social robotics, where the focus is on human-robot interaction (HRI). Social robots are actually autonomous artificial intelligence systems that are designed to interact with humans and other robots. By design they can have many forms – for example, they can resemble people, animals or look like completely new creatures; they can speak and listen, walk on two or four legs (or not move at all); they can communicate in different ways with people and so on. The main purpose of their creation is to be personal assistants, tutors, caregivers, receptionists and many more (Clark & Fisher, 2023). Considering the growing involvement of robots in everyday life, the need to understand and optimize this interaction becomes more and more crucial. Hence, in recent decades, the expansion of such research has been accompanied by the design of social robots that are highly proficient in engaging in joint activities with humans (Belhassein et al., 2022).

As already mentioned, increasing collaboration between humans and robots is a current challenge that interdisciplinary teams of scientists are focusing on. In that direction, the psychology of robotics aims to bridge the gap between humans and robots, providing insight into many specificities of this interaction (Stock & Nguyen, 2019). An innovative interdisciplinary scientific approach between the fields of robotics (social), social psychology (specifically, prosocial behavior) and moral psychology reconstructs the traditional dynamics in human-robot interaction by exploring the ways in which robots seek help from humans and not just focusing on how robot assists humans (which is a one-way street). The uniqueness of this paradigm shift is that it opens new possibilities for increasing the performance of robots and improving the cooperation between them and humans, while in the same time providing an active and mutual process of learning.

[10] https://www.ibm.com/think/topics/artificial-intelligence-types

[11] https://aibusiness.com/verticals/robotics-and-artificial-intelligence-the-role-of-ai-in-robots

Moral reasoning and moral decision-making in different types of moral dilemmas: the role of neuroscience

The field of moral psychology is very complex and interdisciplinary, resembling a puzzle with many pieces, some of them yet to be revealed. It connects interdisciplinary with neuroscience, behavioral economics, sociology, law, anthropology and many more disciplines (Bloom, 2023). Accordingly, morality is one of the most demanding domains for research if one considers the multifactorial conditioned moral behavior. Prosociality can be taken as an example which, according to Decety & Steinbeis (2020), includes many mechanisms based on different motivations and behaviors that in terms of social cohesion and cooperation are significant adaptive elements. The gap between moral knowledge and moral behavior in moral psychology cannot simply be bridged by focusing only on one domain of morality (e.g. moral reasoning, or moral emotions, or moral motivation, etc.). As Green & Haidt (2002) pointed out at the beginning of the 21st century, research from psychology and cognitive neuroscience show that the moral reasoning is many times a matter of affective and intuitive responses and not a result of in-depth and deliberate reasoning. Hence, insights from cognitive and affective neuroscience are very useful, as they provide significant information about brain functioning when the individual is faced with situations that differ in terms of their moral relevance. In this sense, it is already well documented that it really matters whether the individual faces a personal or non-personal, concrete or abstract (hypothetical) moral dilemma, or even a morally neutral dilemma, because when solving different types of dilemmas different brain regions are activated.

When talking about moral reasoning its complexity should be taken into account, given that emotions play a very significant role, as mentioned above. In addition, the human mind has a marked ability to solve problems automatically and unconsciously (which also applies to problems that originate in a complex social context). This is particularly emphasized in the social intuitionism approach which brings together insights from automatic response research, neuroscience and evolutionary psychology. According to this approach, moral judgments are carried out similarly to aesthetic ones, that is, when initially facing a certain moral dilemma the individual intimately feels that the behavior is morally right or wrong, acceptable or unacceptable. Such feelings (which can be positive or negative) appear in consciousness suddenly and without cognitive effort and the individual many times cannot even clearly articulate his/her moral judgment – he/she simply "knows" on a deeper emotional level that something is moral or immoral. The author of this model - Jonathan Haidt - calls them intuitions, and they are created through natural selection and the forces of culture (Green & Haidt, 2002).

Neuroscience research shows that making a moral decision requires interaction between several different yet interconnected brain regions. For example, in two experimental studies Green et al. (2001) used a battery of 60 practical

dilemmas divided into two categories – moral and non-moral. Additionally, moral dilemmas were divided into two groups – personal (emotionally more intense) and non-personal (emotionally less intense). As expected, the findings showed that personal moral dilemmas predominantly engaged brain regions related to emotions (including the MFG, PCG, and ANG - all bilaterally), while non-personal moral dilemmas elicited increased activity in regions associated with cognitive control and working memory (especially in the DLPFC). Another research by Green et al. (2004) showed that the brain regions associated with abstract thinking and cognitive control (primarily rDLPFC and ACC) were activated when a complex (personal) moral dilemma had to be resolved in which the individual, by providing an utilitarian judgment, had to simultaneously violate some of his/hers personal values and principles. Authors also found that parts of the frontal and parietal cortex were more active during utilitarian judgment. On the other hand, the structures of the MPFC were responsible for more intuitive emotional reactions.

According to Green (2023), there is a strong scientific evidence that moral dilemmas elicit competing responses supported by different cognitive systems, with one response described as more emotional and the other as more rational. From the standpoint of the dual-process theory of moral reasoning, the typical deontological judgment of probably the most frequently cited example of a personal moral dilemma - the footbridge dilemma - where the respondent has to answer whether it is acceptable to push a rather large man off the bridge in order to stop the train which is moving on the tracks below him and thus to save five railway workers – that such a thing is completely unacceptable, is supported by an intuitive negative emotional reaction at the very thought of such an action. On the other hand, the typically utilitarian response that it is acceptable to sacrifice one to save five – is supported by a rational assessment on a cost–benefit basis, which respondents understand very well.

However, as Green (2023) further explains, the question of whether deontological response is faster than the utilitarian one is discussed by some researchers since the number of findings showing that the utilitarian response is not so slow is increasing, although a large body of research data really show that it is slower. There are many studies that give clear support to the dual-processing theory of moral reasoning. For example, several research (eg Ciaramelli et al. 2007; Koenigs et al., 2007; Moretto et al., 2010; Thomas et al., 2011; according to Shenhav & Green, 2014) show that patients with impairments in VMPFCs more often make utilitarian judgments. On the other side, those with damage to the hippocampus make deontological judgments significantly more (McCormick, Rosenthal, Miller, & Maguire, 2016; according to Green, 2023), just like patients with damage to the basolateral amygdala important for goal-directed decision-making (van Honk et al., 2022; according to Green, 2023). Such an effect here, as in the case of patients with hippocampus damage, is related to the dominant intuitive emotional reactions that accompany deontological judgments. Hence, according to Green (2023), these two types of judgments – deon-

tological and utilitarian – are driven by different processes. As a matter of fact, according to behavioral data it cannot be said that one process is faster than the other one. Yet, the strong impression remains that deontological judgments are still more intuitive, followed by stronger feelings of right or wrong, while utilitarian ones are visibly more rational.

Research also show that there are potential differences between what one judges to be morally right and how one actually behaves when choosing between alternative actions (i.e. making a moral decision). In that direction, for example, the results of the research of Tassy et al. (2013), point to the conclusion that at the basis of moral reasoning and moral decision-making, different psychological processes are likely to be found. A total of 240 respondents divided into 8 groups in the mentioned research were given 15 moral dilemmas and 9 morally neutral, control dilemmas. For the neutral dilemmas, a large proportion of the respondents gave an appropriate answer, which according to the authors indicates their ability to make appropriate decisions. However, for moral dilemmas, it was observed that in general the decisions about the actions were more utilitarian than the judgments. This basically means that the respondents were inclined to accept to act in a way that they had previously classified as morally unacceptable.

The obtained results from this study also showed that when the number of lives saved was high, a greater number of respondents tended to make utilitarian judgments and choose such behavior. Moreover, the probability of giving utilitarian answers was consistently higher for moral decision, than for moral reasoning. Finally, respondents decisions were less utilitarian when the potential victim was someone who was close to them. This was found in both cases, but the effect was significantly larger for the choice of actions. Moreover, the probability of giving utilitarian answers was higher for choosing an action than for reasoning when potential victim was someone who had low affective closeness to the agent. However, the decisions were completely opposite in a situation of high affective closeness to the potential victim (personal moral dilemma with a high degree of conflict).

Moral psychology of artificial intelligence – is AI capable of making unsupervised moral decisions?

Bonnefon, Rahwan & Shariff (2024) acknowledge that with the development of AI, intelligent machines pop out as a new category moral psychology has to deal with. The attempt to integrate human values into intelligent machines in order to make sound moral judgments and corresponding moral decisions, is still unsuccessful. There is number of research on autonomous vehicles about the possibility of programming them to make a decision whether, when there is a visible obstacle on the road, they will turn towards a group (of five, for example) in order to avoid the obstacle and save the driver's life while kill them, or swerve into a wall (again to avoid the obstacle) and kill the driver. Research

by Bonnefon et al. (as cited in Green, 2016) show that when solving this dilemma respondents generally agree that it is better to save five lives at the expense of one (which is, by the way, a classic utilitarian judgment). But what if the respondent is the driver who is inside the autonomous vehicle? In that case, the respondents change the course of judgment and would not like to drive in such a car. In other words, it is not acceptable for them to die to save five other lives.

Hence, the complexity and importance of integrating morality into engineering is clearly visible. Programmers will face a difficult challenge, as pointed by Green (2016), to integrate morality into algorithms that would program autonomous vehicles to be virtuous and just, taking into account human rights and values. According to this author, such a thing would be possible if there were sufficiently precise moral theories on the basis of which strict criteria and protocols would be developed that will then enable the programmers to know with certainty which virtues these intelligent machines should possess and follow, which human rights exactly should be taken into account when a machine is making moral decision (thus, how the rights are prioritized and whether they are prioritized at all), as well as how these machines can become able to make fair compromises.

McKendrick & Thurai (2022)[12] in the article "*AI is not ready to make unsupervised decisions*", published in the Harvard Business Review, write that AI is designed to help making decisions when there are various types of data that surpass human understanding. However, the authors emphasize that AI is unable to address all relevant human factors that are involved into real-life decision-making. Pazzanese (2020)[13] in the review "*Great promise, but potential danger*", published in The Harvard Gazette, emphasizes that AI is starting to be used more intensively in many fields thus increasing ethical concerns related to its (in)ability to make moral decisions. It is already implemented, for example, in medicine (for making a diagnosis, or for estimating which patient should have priority for receiving an organ transplant), in banking (for coming to a decision who should be given credit), in the military industry (for production of autonomous lethal weapons), in the judiciary (for deciding who should go to prison) etc.

Research by Zhang, Chen, & Xu (2022) regarding the perception of moral decision-making by AI, which presented findings from 4 experimental studies and a total of 804 respondents, reported that when compared to humans, respondents perceived AI as more inclined to make utilitarian decisions in moral dilemmas. They also perceive AI as more competent than humans, but at the same time less emotionally "warm". Another insight is that individuals may behave less morally and may be more willing to deceive others when communicating through AI. The research by Zhang et al. (2023) on human versus AI

[12] https://hbr.org/2022/09/ai-isnt-ready-to-make-unsupervised-decisions

[13] https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role

moral judgments in moral dilemmas (the footbridge dilemma and the trolley dilemma), which reported results from three experiments and 626 respondents, showed that in the trolley dilemma, respondents rated AI as more immoral and considered that its decisions deserve more condemnation compared to those of humans. On the other side, in the footbridge dilemma, respondents rated utilitarian behavior (regardless of the entity) as more immoral thus being less permissive, and more wrong than deontological behavior (taking no initiative). The bottom line is that in different types of moral dilemmas, people apply different types of moral reasoning when evaluating AI's behavior.

The results of the experimental research of Tolmeijer et al. (2022), with a total of 428 respondents, where was considered how the level of expertise (human vs. AI) and the level of expert autonomy (advisor vs. decision maker) affects trust, perceived responsibility and reliance on the source, show that the respondents perceived the human- an expert as more trustworthy (from a moral point of view), but less capable than an AI. Moreover, respondents more often accepted the recommendations and decisions of the AI-expert than the human-expert, while they perceived the AI-expert as less responsible than humans.

Bonnefon, Rahwan & Shariff (2024), refer to intelligent machines as moral actors both implicitly and explicitly, depending on the purpose for which they are programmed. When they talk about implicit moral machines, the authors mean machines that if they make a mistake can cause harm to someone, even though they are not initially programmed for encoding moral values (e.g. if they mistake a medical diagnosis or misrecognizing someone as a wanted criminal). Explicit moral machines, on the other hand, are either programmed to solve moral dilemmas (e.g. when estimating which patient from the list should have priority for organ transplantation), or there is a possibility of facing a moral dilemma in a certain situation, which is why they should be able to deal with it successfully (e.g. while favoring speed and precision in work, on the expenses of empathy and long-term psychological well-being of employees in the organization).

In addition to all of this, according to these authors, there are many unanswered questions. For example, the questions of how many errors can be allowed to the AI and what the nature of those errors should be? Or, how would people react to AI errors? How much would they blame the AI when it makes an error on its own compared to when it shares the responsibility for the error with a human? Should humanity perhaps wait until a version of a perfect intelligent machine that make 0% errors is built, before putting it into practice? At the same time, doesn't it mean that a possibility for saving many peoples' lives, for example by early diagnostics, would be jeopardized? And this is not all. There are many other known and unknown questions and dilemmas. For some of them, there are already recent studies that show that when an implicit moral machine makes a mistake, people have much stronger negative reactions (because they have very high expectations from it), than when a human

makes a similar mistake. This is clearly the case in research on autonomous vehicles, where respondents are much more inclined to condemn a traffic accident caused by an autonomous vehicle as more serious and less acceptable, assigning much more responsibility and blame to it, than if the same accident were caused by a human (e.g. Franklin et al., 2021; Hong et al., 2022; as cited in Bonnefon, Rahwan & Shariff, 2024). However, the results change when a human is involved in the task performed by the intelligent machine. When an autonomous vehicle and its driver are involved in car accident (e.g. hit a pedestrian), respondents are more likely to blame the driver. This difference in findings is not yet clear enough and continues to be investigated.

Hence, the answer to the question of whether AI can make unsupervised moral decisions is: no, not yet. This conclusion follows from everything stated so far. Considering the complexity of moral reasoning and decision-making processes, as well as the fact that moral dilemmas that can be of different types, contain a conflict between two moral values which intelligent machines are incapable to resolve; then that solving these dilemmas, simultaneously involves both cognitive and affective processes, but also automatic and intuitive responses (which is again absent in AI), that activates different brain regions and neural networks; but also that the artificial neural networks generative AI tools rely on are not yet as complex as the human brain (although they try to simulate its activity); as well as that moral reasoning relies to a large extent on specific processes of social cognition and on the representation of the mental states of other people, i.e. the theory of mind which, along with emotional intelligence and self-awareness, is unattainable for AI, it is clear that there are still many unresolved issues and steps to climb before AI becomes capable of making moral decision independently, with 0% of error.

Finally, it is more than obvious that humanity is facing a serious challenge to respond to the progress of AI and its collision with, as Bonnefon, Rahwan & Shariff (2024) will write "...the moral intuitions of people forged by culture and evolution over the span of millennia." (p.669). It requires great knowledge, great ability to adapt, open-mindedness, visionary, cooperation, conscientiousness and responsibility from all relevant actors involved in the process of designing and programming algorithms, autonomous systems as well as in integrating morality into intelligent machines. How far the humanity will go in that regard and whether one day advanced autonomous systems with the same intelligence as that of humans will "conquer" the world, time will tell. However, as long as there is order and control over algorithms, and AI is consciously and conscientiously used as a tool to improve people's quality of life, things will not get out of hand. In that direction, specific steps have already been taken by the European Union with the adoption of the EU AI Act of April 19, 2024, which introduces new legislation and lays the foundations for AI use within the borders of the Union.

BIBLIOGRAPHY:

- Belhassein, K., Fernandez-Castro, V., Mayima, A., Clodic, A., Pacherie, E., Guidetti, M., Alami, R., Cochet, H. (2022). Addressing joint action challenges in HRI: Insights from psychology and philosophy. *Acta Psychologica* 222, pp. 1-14 Достапно на: https://doi.org/10.1016/j.actpsy.2021.103476
- Berruti, F., Nel, P., Whiteman, R. (29 April, 2020). *An Executive Primer on Artificial General Intelligence*. McKinsey & Company. Достапно на: https://www.mckinsey.com/capabilities/operations/our-insights/an-executive-primer-on-artificial-general-intelligence (Пристапено на 20.06.2024)
- Bloom, P. (2023). *The Human Mind. A Brief Tour of Everything We Know*. Penguin Random House, UK
- Bonnefon, J.F., Rahwan, I., Shariff, A. (2024). The Moral Psychology of Artificial Intelligence. *Annual Review of Psychology*, 75, pp. 653-75 https://doi.org/10.1146/annurev-psych-030123-113559
- Clark, H.H. & Fisher, K. (2023). Social robots as depictions of social agents. *Behavioral and Brain Sciences* 46, e21, pp. 1–65 Достапно на: https://doi.org/10.1017/S0140525X22000668
- Decety, J., Steinbeis, N. (2020). Multiple Mechanisms of Prosocial Development. In: Decety, J. (ed.) (2020). *The Social Brain. A Developmental Perspective*, The MIT Press, pp. 219-246
- Green, J.D. (2023). Dual-process moral judgment beyond fast and slow. Commentary. *Behavioral and Brain Sciences*, 46 e123, pp. 35-36 https://doi.org/10.1017/S0140525X22003193
- Green, J.D. (2016). Our driverless dilemma: When should your car be willing to kill you? *Science 352*, pp. 1514-1515 doi: 10.1126/science.aaf9534
- Green, J.D., Haidt, J. (2002). How (and where) does moral judgment work? *TRENDS in Cognitive Sciences* 6 (2), pp. 517-523 https://doi.org/10.1016/S1364-6613(02)02011-9
- Green, J.D., Nystrom, L.E., Engell, A.D., Darley, J.M., Cohen, J.D. (2004). The Neural Bases of Cognitive Conflict and Control in Moral Judgment. *Neuron*, 44 (2), pp. 389-400 https://doi.org/10.1016/j.neuron.2004.09.027
- Green, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M., Cohen, J.D. (2001). An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science* 293, pp. 2105-2108 doi: 10.1126/science.1062872
- Hatzius, J., Briggs, J., Kodnani, D., Pierdomenico, G. (26 March, 2023). The Potentially Large Effects of Artificial Intelligence on Economic Growth (Briggs/Kodnani). *Economics Research*, Global Economics Analyst. 1-20 Достапно на: https://www.key4biz.it/wp-content/uploads/2023/03/

Global-Economics-Analyst_-The-Potentially-Large-Effects-of-Artificial-Intelligence-on-Economic-Growth-Briggs_Kodnani.pdf (Пристапено на 20.06.2024)

- Kelly, J. (28 February, 2024). *What White-Collar Jobs Are Safe From AI—And Which Professions Are Most At Risk?* Forbes. Достапно на: https://www.forbes.com/sites/jackkelly/2024/02/28/what-white-collar-jobs-are-safe-from-ai-and-which-professions-are-most-at-risk/ (Пристапено на 20.06.2024)
- Martin, A. (26 November, 2021). *Robotics and Artificial Intelligence: The Role of AI in Robots*. AI Business. Достапно на: https://aibusiness.com/verticals/robotics-and-artificial-intelligence-the-role-of-ai-in-robots (Пристапено на 20.06.2024)
- McKendrick, J., Thurai, A. (15 September, 2022). *AI Isn't Ready to Make Unsupervised Decision*, Harvard Business Review, Достапно на: https://hbr.org/2022/09/ai-isnt-ready-to-make-unsupervised-decisions (Пристапено на 20.06.2024)
- Pazzanese, Ch. (26 Ocober, 2020). *Great Promise but Potential for Peril*, The Harvard Gazette, Достапно на: https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/ (Пристапено на 20.06.2024)
- Shenhav, A., Green, J.D. (2014). Integrative Moral Judgment: Dissociating the Roles of the Amygdala and Ventromedial Prefrontal Cortex. *The Journal of Neuroscience*, 34 (13), pp. 4741-4749 DOI: 10.1523/JNEUROSCI.3390-13.2014
- Stock, R.M. & Nguyen, M.A. (2019). Robotic Psychology: What Do We Know about Human–Robot Interaction and What Do We Still Need to Learn? *Proceedings of the 52nd Hawaii International Conference on System Sciences*, pp. 1936-1945 Достапно на: https://scholarspace.manoa.hawaii.edu/items/7ab2be7a-3a3a-463e-89b3-9753041f7e19
- Tassy, S., Oullier, O., Mancini, J., & Wicker, B. (2013). Discrepancies between judgment and choice of action in moral dilemmas. *Frontiers in Psychology*, 4, pp. 1-8 https://doi.org/10.3389/fpsyg.2013.00250
- Tolmeijer, S., Christen, M., Kandul, S., Kneer, M., & Bernstein, A. (2022). Capable but Amoral? Comparing AI and Human Expert Collaboration in Ethical Decision Making. In: *ACM CHI Conference on Human Factors in Computing Science (CHI'22)*, New Orleans, LA, USA, 29, April 2022-5 May 2022, ACM Press https://doi.org/10.1145/3491102.3517732
- Zaixuan, Z., Chen, Z., & Xu, L. (2022). Artificial intelligence and moral dilemmas: Perception of ethical decision-making in AI. *Journal of Experimental Social Psychology*, 101, pp. 1-8 DOI:10.1016/j.jesp.2022.104327

- Zhang, Y., Wu, J., Yu, F., & Xu, L. (2023). Moral Judgments of Human vs. AI Agents in Moral Dilemmas. *Behavioral Science*, 13, 181, pp. 1-14 https://doi.org/10.3390/bs13020181

Consulted relevant professional web-sites:

- https://www.ibm.com/think/topics/artificial-intelligence-types (Accessed on 10.06.2024)
- https://www.resumebuilder.com/1-in-3-companies-will-replace-employees-with-ai-in-2024/ (Accessed on 10.06.2024)
- https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-artificial-general-intelligence-agi#/ (Accessed on 10.06.2024)
- https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai (Accessed on 10.06.2024)
- https://artificialintelligenceact.eu/ai-act-explorer/ (Accessed on 20.06.2024)