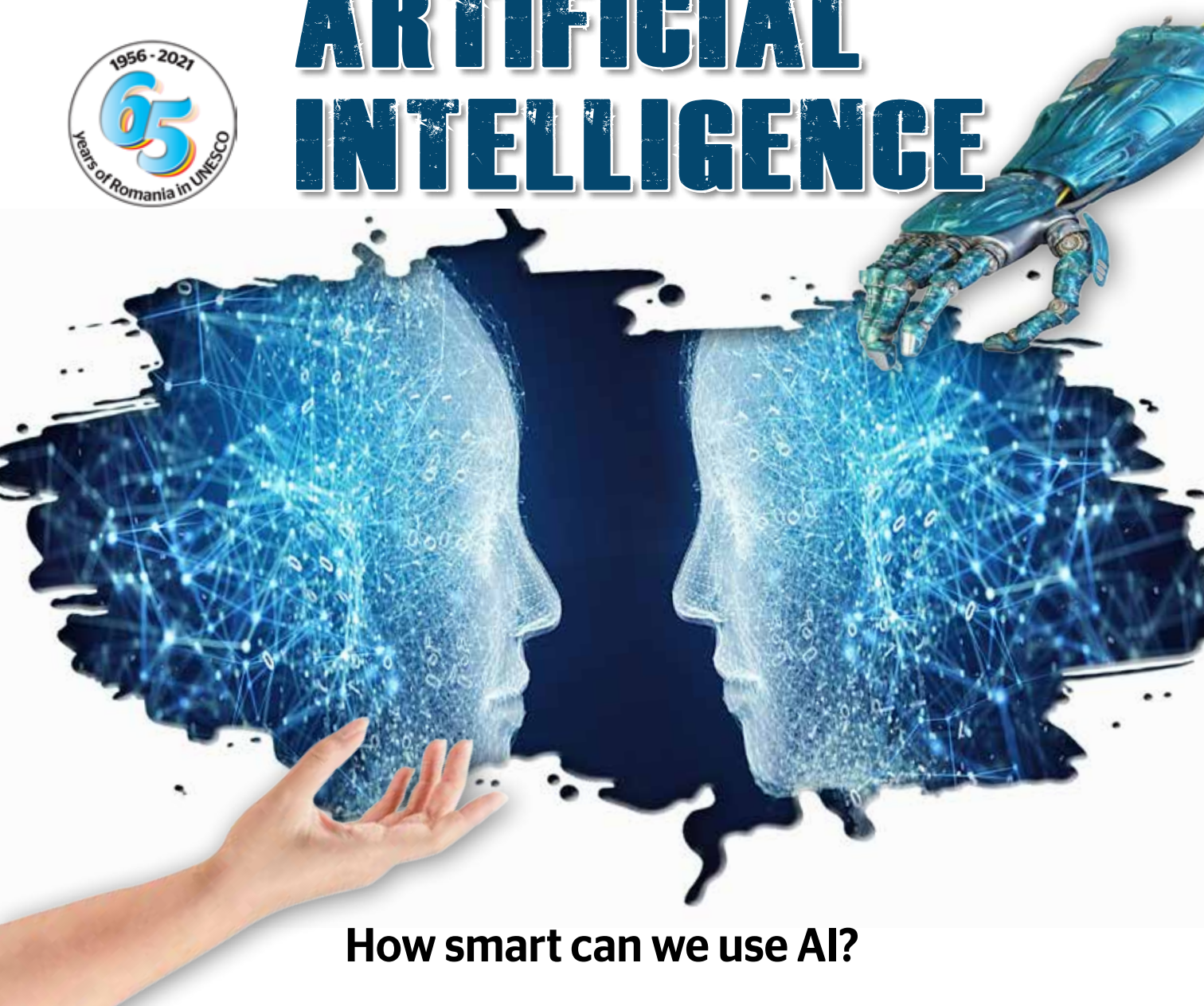
unesco
National Commission
of Romania for UNESCO

1956 - 2021
65
Years of Romania in UNESCO

ETHICS ON ARTIFICIAL INTELLIGENCE



How smart can we use AI?

unesco
National Commission
of Romania for UNESCO

NATIONAL COMMISSION OF ROMANIA FOR UNESCO

Str. Anton Cehov nr. 8, sector 1, București, 011998
Tel.: +4 (021) 231.13.33/231.32.24
Fax: +4 (021) 230.76.36
cnr@cnr-unesco.ro
www.cnr-unesco.ro

**„Ethics of Artificial Intelligence.
How smart can we use AI?“**

**This publication is part
of the anniversary program
“Romania and UNESCO, bridges over time -
65 years of NRC UNESCO “**

Thanks to the International Center of Excellence in Artificial Intelligence at University POLITEHNICA of Bucharest for the overstanding contribution.

Photo Credit:SHUTTERSTOCK

Ethics of Artificial Intelligence. How smart can we use AI?

Artificial Intelligence will redefine new territories in Science

Madlen Şerban
Secretary General
National Commission of Romania for UNESCO

Three years ago, Audrey Azoulay, Director-General of UNESCO, launched an ambitious project: to give the world an ethical framework for the use of artificial intelligence.
In March 2020, the Director-General appointed an ad-hoc multi-disciplinary group of 24 specialists to produce a draft text of a UNESCO Recommendation, taking into account the various contemporary global issues.
Following the evolution of the draft Recommendation, the National Commission of Romania for UNESCO aligned itself with the proposed goal and objectives of the draft, trying to organise, on a smaller scale, the debates on such a challenging topic as Ethics of Artificial Intelligence.
Thus, on June 3, 2021, CNR UNESCO organized the online colloquium "Ethics of Artificial Intelligence. How smart can we use AI?", which was attended by members of the Romanian scientific community in the field and by a foreign guest from the French National Commission for UNESCO.
On October 7, during the event "UNESCO's Days at the POLITEHNICA", the UNESCO chair hosted another colloquium on the same topic.
This publication authored by a group of professors and researchers in the field of ethics and AI, concludes the series of projects dedicated to Ethics of Artificial Intelligence, without exhausting, in any way, the myriad of the subject perspectives. This publication is about intellectual and moral solidarity!
It is about to be now and in the future that began yesterday. It is the essence of being of UNESCO, which supports, together with other international organizations, member states, associated and other states, the development of new technologies, through cooperation and transparency. Artificial Intelligence, like any technology, offers many benefits. I acknowledge and express my appreciation to Florina Pescaru, NCR UNESCO expert for coordinating the entire project!
The distinguished authors of this publication highlighted the AI applicability in various fields. But, unfortunately, technology can also be used against humanity or misused.
Through the authors contribution, we learn about explainable, comprehensible and trustworthy AI.
What are the challenges of using AI in medicine – for example, if algorithms interpreting thoracic radiographs are trained with data from mainly male patients, the results are not as accurate when applied to interpreting the thoracic radiographs of female patients. What are the AI applications in the aerospace industry and airports? What is ethical and unethical in AI or how to use the new technologies for the automatic detection, inter alia, of unethical posts or messages, bullying, misogyny, xenophobia.
Some articles also try to answer questions such as: How can we ensure that ethical recommendations are reflected in the practice of implementing AI in different markets? What are the harmful side effects of AI implementation in terms of moral externalities, not economic externalities?
Some authors have emphasized that legislating morality is not a simple task and remains one of the important issues in moral philosophy, political philosophy and the philosophy of law. The adoption of the UNESCO's Recommendation for the Ethics of AI is an important step to find, along the way, the multiple answers to these and other questions to come.
We are at the beginning of the journey. We know that it is not just a journey, but a knowledge discovery and development Way, which we will travel together, in this New Interconnected World.

The content of the Recommendation on the Ethics of AI. Key points

The Recommendation aims to realize the advantages AI brings to society and reduce the risks it entails. It ensures that digital transformations promote human rights and contribute to the achievement of the Sustainable Development Goals, addressing issues around transparency, accountability and privacy, with action-oriented policy chapters on data governance, education, culture, labour, healthcare and the economy.

1. Protecting data. The Recommendation calls for action beyond what tech firms and governments are doing to guarantee individuals more protection by ensuring transparency, agency and control over their data. It states that individuals should all be able to access or even erase records of their data. It also includes actions to improve data protection and an individual's knowledge of, and right to control, their own data. It also increases the ability of regulatory bodies around the world to enforce this.

2. Banning social scoring and mass surveillance. The Recommendation explicitly bans the use of AI systems for social scoring and mass surveillance. These types of technologies are very invasive, they infringe on human rights and fundamental freedoms, and they are used in a broad way. The Recommendation stresses that when developing regulatory frameworks, Member States should consider that ultimate responsibility and accountability must always lie with humans and that AI technologies should not be given legal personality themselves.

3. Helping to monitor and evaluate. The Recommendation also sets the ground for tools that will assist in its implementation. Ethical Impact Assessment is intended to help countries and companies developing and deploying AI systems to assess the impact of those systems on individuals, society and the environment. Readiness Assessment Methodology helps the Member States to assess how ready they are in terms of legal and technical infrastructure. This tool will assist in enhancing the institutional capacity of countries and recommend appropriate measures to be taken in order to ensure that ethics are implemented in practice. In addition, the Recommendation encourages Member States to consider adding the role of an independent AI Ethics Officer or some other mechanism to oversee auditing and continuous monitoring efforts.

4. Protecting the environment. The Recommendation emphasizes that AI actors should favour data, energy and resource-efficient AI methods that will help ensure that AI becomes a more prominent tool in the fight against climate change and on tackling environmental issues. The Recommendation asks governments to assess the direct and indirect environmental impact throughout the AI system life cycle. This includes its carbon footprint, energy consumption and the environmental impact of raw material extraction for supporting the manufacturing of AI technologies. It also aims at reducing the environmental impact of AI systems and data infrastructures. It incentivizes governments to invest in green tech, and if there is a disproportionated negative impact of AI systems on the environment, the Recommendation instructs that they should not be used.

Emerging technologies such as AI have proven their immense capacity to deliver for good. However, its negative impacts that are exacerbating an already divided and unequal world, should be controlled. AI developments should abide by the rule of law, avoiding harm, and ensuring that when harm happens, accountability and redressal mechanisms are at hand for those affected.

Source: https://en.unesco.org/news/unesco-member-states-adopt-first-ever-global-agreement-ethics-artificial-intelligence

Artificial Intelligence for a better life

Florina Pescaru
Expert, Science Subcommittee,
National Commission of Romania for UNESCO

Today, Artificial Intelligence plays a major role in billions of people's lives on Earth. Sometimes unobserved in laboratories, on the Internet, on orbital research stations, but often with profound consequences, it transforms our societies and changes the quality of our life.

During the Covid-19 pandemic, when online education became a necessity, the immersion of AI in our existence accelerated. In recent years, the safety of the growing volume of banking transactions has been possible thanks to algorithms.
The explosion of information (big data) and the accelerated use of AI will energize more and more all sectors of our societies: we will have better medical services, safer vehicles and transportation systems, cheaper and more sustainable services and products.
AI can facilitate access to education and training and improve our safety in the workplace by robots taking on the tasks considered dangerous.
AI is helping companies to carry out remote working and management improving operational efficiency.
Most major healthcare organizations are relying on AI-based software for their everyday tasks.
Every day the media reports the news about major discoveries in science such as by using a machine-learning algorithm the researchers have identified a powerful new antibiotic compound or they will use AI to discover planets outside our solar system.
The processes are so disruptive that existential fears arise: where does the good end and where does the evil produced by AI begin?
In his recently published book, „The Age of AI and Our Human Future", the former statesman Henry Kissinger said: „Artificial intelligence could be as important as the advent of nuclear weapons, but less predictable".
„Decisions impacting millions of people should be fair, transparent and contestable. These new technologies must help us address the major challenges in our world today, such as increased inequalities and the environmental crisis, and not deepening them", said Gabriela Ramos, UNESCO's Assistant Director-General for Social and Human Science.
UNESCO considers that global reflection on AI is of crucial importance to ensure that new technologies, notably those based on AI, serve the good of societies, contribute to sustainable development and respect human rights and human dignity.
In this spirit, on 24 November 2021, UNESCO has adopted a comprehensive global standard-setting instrument to provide AI with a strong ethical basis. It will not only protect but also promote human rights and human dignity, and will be an ethical guiding compass and a global normative bedrock allowing to build strong respect for the rule of law in the digital world.

Ethical Principles in Artificial Intelligence

Prof. Dr. Eng. Adina Magda Florea,
University POLITEHNICA of Bucharest

Artificial Intelligence (AI) aims to build systems that exhibit rational behavior, analyze the environment and make autonomous decisions to perform specific tasks. AI-based applications are ubiquitous today, from smartphones to robots, autonomous machines and smart assistants, to machine translation, synthesizing opinions from huge volumes of text and social media posts, to name a few. We can already say that we are in the age of Artificial Intelligence and this is just the beginning. However, integrating AI-based decision-making systems into everyday life, no matter how rational their behavior, is not straightforward for many reasons.

There is a number of important questions that have to be asked about our future when AI technology and applications will be ubiquitous. Will Artificial Intelligence become more „intelligent" than human intelligence? Is AI creating welfare and opportunities for all? Are we, humans, to trust AI autonomous decisions and are we endangered in any way by these decisions? While the issue of whether or not AI will become more intelligent than humans is still open to high debates, the other questions, and many related ones, are now among the central concerns of international and national associations and organizations, high tech companies and civil society as well.

UNESCO has stated that „*we need a human-centered AI, which must be for the greater interest of the people, not the other way around*" and proposed the development of a comprehensive global standard-setting instrument to provide AI with a strong ethical basis, that will not only protect but also promote human rights and human dignity. UNESCO published the „*Draft text of the Recommendation on the Ethics of Artificial Intelligence*"[1] in June 2021, which was submitted to the General Conference at its 41st session for adoption. The recommendation coagulated 10 principles for the realization of ethical AI:

- **Proportionality and Do No Harm** – AI use must be proportional to achieve a given legitimate aim and must not violate human rights;
- **Safety and security** – avoid unwanted harms and vulnerabilities to attack;
- **Fairness and non-discrimination** – ensure that the benefits of AI are available and accessible to all, minimize and avoid reinforcing or perpetuating discriminatory or biased applications;
- **Sustainability** – continuous assessment of the human, social, cultural, economic and environmental impact of AI technologies;
- **Right to Privacy and Data Protection** – adequate data protection frameworks and governance mechanisms;
- **Human oversight and determination** – an AI system can never replace ultimate human responsibility and accountability;
- **Transparency and explainability** – the capacity

[1] https://unesdoc.unesco.org/ark:/48223/pf0000377897

of an AI system to explain its decision process;

- **Responsibility and accountability** – respect for human rights and fundamental freedoms, protection of the environment, auditability and traceability of AI systems;
- **Awareness and literacy** – raise public awareness and understanding of AI technologies, make data open and AI education accessible for all;
- **Multi-stakeholder and adaptive governance and collaboration** – large participation, governmental and international regulations for both AI and data use.

The strategy on Artificial Intelligence of the European Commission aims to „*develop and deploy cutting-edge, ethical and secure AI, promoting a human-centric approach in the global context*". The High-Level Expert Group on Artificial Intelligence[2] supporting the implementation of the European Strategy on AI identified four ethical principles for the foundation of trustworthy AI: *Respect for human autonomy, Prevention of harm, Fairness, and Explicability*, and seven key requirements for the realization of trustworthy AI: human involvement and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; environmental and societal well-being; and accountability.

To apply these principles in real life, different stakeholders throughout the entire AI system life cycle must participate: researchers and academia, the technical community, private sector companies, governments, intergovernmental organizations, and the civil society. Artificial Intelligence must be in support of humans and their fundamental rights, and we all have the very challenging task to further develop and deploy AI technologies and systems in responsible and robust ways, set up rules and regulations that are politically, legally and ethically acceptable, for the benefits of all people and for the realization of a sustainable and flourishing society.

Photo Credit: PIXABAY

[2] https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

AI charts its role in airport operations and passenger experience

Prof. Dr. Eng. Sorin Eugen Zaharia,
University POLITEHNICA of Bucharest

Today's air transport is indispensable to modern life, and it is a necessity for economic prosperity in a globalized economy. The following happened daily in 2019: 12.5 million passengers traveled by plane; 128,000 scheduled flights were made; $ 18 billion worth of goods were transported[1].

Airports are constantly launching new digital initiatives and artificial intelligence (AI) can be a great asset allowing air transport to optimize operations and helping aviation to continuously improve safety and security, two of the most needed priorities for both airports and air transport providers. Airports are launching dedicated apps to cover key areas where digitalization has a huge impact: operations, security, collaborative decision making, predictive & preventive solutions, customer engagement and retail. Digital solutions have also played a significant role in regaining passenger's trust and in traffic recovery during the pandemic.

Digital transformation is mainly about using technologies in the automation process and passenger engagement, which involve mobile Customer Relationship Management, cloud, block-chain technologies, big data, Internet of Things (IoT), or robotics. Another important aspect refers to flow monitoring which applies predictive/preventive solutions to airport indoor geolocation, identity management, flow management or radio frequency identification (RFID). Concerning the efficiency of airport operations, a harmonized approach between different actors is important. It involves cognitive systems based on the analysis of integrated data from appropriate process monitoring systems in order to predict and improve airport processes.

Managing an airport requires a close collaboration of all actors for the optimal allocation of resources via digital means. The influence of organizational culture in modern aeronautical organizations may represent a key to the success of the digital transformation. Based on AI, airport designers use many ways, presented below, in which they try to find the best solutions for enhancing operations and passenger experience.

A self-driving robot experimented at Munich, Frankfurt and Changi airports is an AI-based transport and delivery robot that accompanied transit passengers to their gates and helped them trans-

[1] ATAG, Aviation benefits beyond borders, 2021

port small luggage thanks to its integrated navigation system. How AI is used is often invisible to the passenger. Behind the scenes, airports management is planning to introduce AI models **to calculate passenger forecasts**. These models will consider all available data – from weather to traffic to passenger numbers – and predict how many people will be in any given location at a specific time. The airports can be better prepared for passenger peaks and adjust staff allocation accordingly or divert them to other process points by calculating the time and taking passengers to make their journey through the airport and logging their progress at every step of the way.

Removing aircraft from parking lots, another AI-based project, could improve the forecast time for ground operations. The method is to calculate, together with external partners, the timing of ground operations of hundreds of thousands of flights and to identify key factors that could affect the prediction of operations on the platform. Video analysis used to support prediction can turn standard CCTV systems into intelligent and efficient detection and alert systems.

By providing automating services and using big data, Copenhagen Airport has enhanced efficiency and became one of the world airports that applies most self-service solutions or utilizes technologies to reduce waiting times in the baggage reclaim or speed up boarding times[2]. The winner of the digital transformation award in 2017, Singapore Changi Airport has managed to create **a unified airport identity by digitalizing core processes and operations** and using data platforms for problem-solving and collaboration between business partners. From a **total airport management** perspective, Changi has reached better operational anticipation and reaction and has improved resource planning, acquiring a unified digital identity[3].

Despite the lengthy process going into developing a powerful AI model, the benefits heavily outweigh the costs. AI can be a great asset, allowing airports to relieve existing personnel of the burden of mundane and repetitive tasks, having more time to take on the more specialized and fulfilling ones[4].

The AI applied by airports is only as reliable as the data that it uses. Therefore, their first challenge is gathering that data, ensuring that there is enough of it and that it is high quality, especially for very big airports which work with hundreds of companies. After overcoming these, the implementation of AI needs to make sure that the data is being correctly interpreted and any data processed is safe, secure and GDPR compliant.

When the AI is implemented on the airport, a question may arise: Are staff and passengers wary of AI? Whether it is increased punctuality from AI technology in the tower, which could help reduce delays or more accurate passenger forecasts that will help predict when the security lanes will be at their busiest, the technologies benefit both passengers and staff equally.

[2] Kobenhavns Lufthavne, Europe's most efficient airport is in Copenhagen, 2016, https://www.cph.dk/en/ news/2016/6/europes-most-efficient-airport-is-in-copenhagen
[3] Jimenez, D. Z., Afuang, A., Rago, T., Tew,, K. L., Changi Airport Group wins Digital Transformer of the Year for Singapore at the 2017 IDC Digital Transformation Awards (DXa), Sep 2017, ttps://www.idc.com/getdoc.jsp?containerId=prAP43099417
[4] International Airport Review, Issue 01, February 2020

The purposeful and ethical use of AI can transform the air transport industry and solve several business challenges in a way that hasn't previously been possible. The more efficient operation of airports benefits passengers, airlines, personnel and local communities. As with the use of any technology, the airports monitor how the progress in AI applications is doing and what the wide societal reaction to it is. This is just the start, and air transport is looking forward to working with all the stakeholders to realize its full potential. Being on this journey already puts airports in an inspiring position.

Photo Credit: PIXABAY

Ethics of Artificial Intelligence for Health

Professor PhD. Alexandru Scafa-Udriște,
University of Medicine and Pharmacy Carol Davila

Professor PhD. Adina Magda Florea,
University POLITEHNICA of Bucharest

Artificial Intelligence (AI) is transforming every aspect of modern society and is likely to have an enormous impact in the coming decades. AI technologies may help promote inclusive economic growth, bring great benefits to society, and empower individuals. Furthermore, AI demonstrates a high potential in contributing to solve global challenges, such as improved medical care, including the fight against global pandemics. AI has the potential to improve personal and global health, to reduce disparities between health systems in different countries and to contribute to personalized healthcare. Nevertheless, the use of AI tools and applications in medicine raises new challenges from the point of view of the ethical implications of its deployment, for example, the data used to train these applications are prone to introduce new kinds of errors. Therefore, AI applications might create risks, lead to unintended harm and challenges in relation to legal liability and responsibility, while undermining human rights and due processes.

Several aspects of using AI for health have been put forth by different organizations, for example the recent "Ethics & Governance of Artificial Intelligence for Health"[1] issued in June 2021 by the World Health Organziation, which identifies consensus principles to ensure that AI works to the public benefit of all countries and a set of recommendations in using AI for health.

The ethical issues surrounding AI in the field of health are complex and refer to different aspects of protecting human rights and ensuring individual well-being. A human-centered perspective of AI for health means that humans should remain in full control of healthcare systems and medical decisions. Moreover, the use of medical data for training AI systems must protect the privacy and con-

[1] https://www.who.int/publications/i/item/9789240029200

fidentiality of patients and ensure informed, valid consent by adopting appropriate legal frameworks for data protection. For example, according to a recent study, physicians devote 62% of their time per patient reviewing electronic health records (EHRs), with the most time-consuming portion being clinical data review. AI systems to assist physicians in reviewing EHRs can increase efficiency and allow them to invest more time in patient care. However, the massive volumes of data needed by machine learning systems may be used for purposes for which the data was not initially intended, data can be re-identified even if anonymized, and has the potential to be hacked and used for harmful or commercial aims.

Another important ethical issue related to how decisions are made by an AI health system is bias in data, in case the training data is incomplete, inaccurate, or unrepresentative for different populations, and bias in algorithms, if developed by humans with implicit bias. Bias may potentially lead to perpetuating or even amplifying inequalities based on race, gender identity, age or demographic characteristics and limit the performance of diagnosis or treatment decisions. For example, if algorithms for reading chest X-rays are trained with data from primarily male patients, the results are not as accurate when applied to chest X-rays of female patients. Skin-cancer detection algorithms trained primarily on light-skinned individuals do worse at detecting skin cancer affecting darker skin. An AI system for recognizing human activities, when applied to support elderly people in recovering from a disease, may perform poorly if trained only for activities done by young persons.

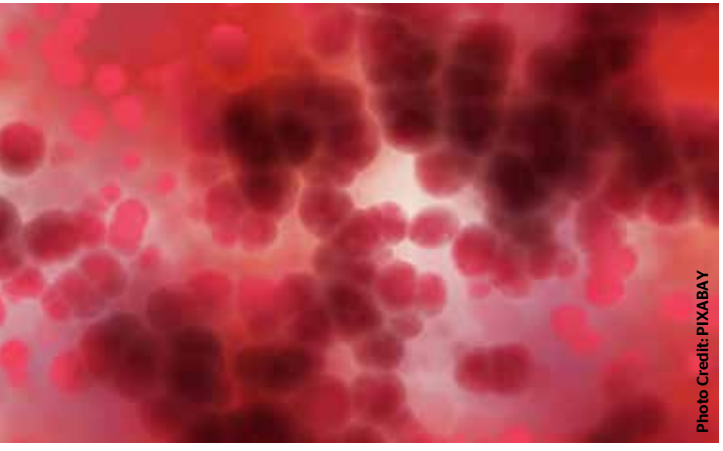


Photo Credit: PIXABAY

Accountability and responsibility are, as well, important ethical issues when using AI for health: who ought to assume responsibility for error of clinical diagnosis and treatment? This is a difficult question as many AI systems are opaque, explainability and transparency being a current challenge in AI research. Assistive robots for taking care of elderly are becoming more and more used as humanoid like robots advance in performance and decrease in price. However, if such a robot harms a person, who is going to be responsible for that harm: the designer of the mechanical part, the software engineer or the medical doctor who recommended the tasks undertaken by the robot? The use of AI technologies must not result in any mental or physical harm.

All these questions and ethical challenges must be discussed and properly answered as AI for health has an immense potential but must not undermine human rights and equal access to health care.

Ethics in Artificial Intelligence

Prof. Dr. Eng. Ștefan Trăușan-Matu,
University POLITEHNICA of Bucharest

Artificial intelligence (AI) is a highly interdisciplinary field, which is now ubiquitous in everyone's life, with major effects on society and individuals. Therefore, ensuring compliance with ethical principles is a very important issue. For example, there are many applications with AI that can generate ethical issues: conversational agents, facial recognition, extracting knowledge from conversations or from posts via email or social networks, assisting robots for the elderly or disabled, autonomous vehicles, etc. In these applications you can encounter several types of ethical issues that can affect people: generating unethical remarks in conversations, bias, making wrong decisions, using the knowledge extracted from posts on social networks for unethical purposes. For example, there are known cases where programs trained with AI techniques have been biased, for example, in granting bank loans or conditional release from prison (according to the White Paper on Artificial Intelligence[1]). User profiles of a social network, built with AI techniques can be used for unethical purposes. On the other hand, the artificial intelligence of an autonomous car, what decision to make in the event of an imminent accident that will affect more people: who should be injured and who should not?

In order to avoid situations where ethics are violated in the context of AI, documents of the European Union, UNESCO, the Council of Europe[2], papers of several researchers and first-class teachers in the world have been published. For example, in the White Paper on Artificial Intelligence an important part is dedicated to ethics. The European Parliament has also published a document highlighting the need for a human-centric AI approach[3]. On the other hand, within large companies, such as IBM or Orange, doc-

[1] https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf
[2] https://rm.coe.int/algorithms-and-human-rights-en-rev/16807956b5
[3] https://www.europarl.europa.eu/RegData/etudes/BRIE/2019/640163/EPRS_BRI(2019)640163_EN.pdf

Photo Credit: PIXABAY

uments have been developed to analyze the issue of ethics in the context of AI.

The European Commission's AI HLEG Expert Group has published an "Assessment List for Trustworthy Artificial Intelligence" (ALTAI)[4]. In this document are highlighted several dangers that can arise in the context of the explosion of applications using artificial intelligence. Moreover, actions to be taken and the subjects and entities to be considered were also identified. The HLEG AI Expert Group has identified 7 essential requirements for the development of AI: human involvement and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; environmental and societal well-being; and accountability.

The investigations on the ethical aspects of AI should answer a few questions: What are the ethical implications in the use of AI technology? How can we verify the fulfillment of the ethical issues? What are the possibilities of implementing for robots, agents or other AI programs that take into account implicit or explicit ethical principles?

Detecting situations of an ethics violation in the AI context has two aspects: 1) Avoiding generation of unethical situations due to AI, 2) The usage of AI technologies for the detection of ethical violations by other agents, human or artificial, such as the automatic detection of unethical posts or messages, for example, bullying, misogyny, xenophobia, etc. The implementation of the detection of violation situations of ethics is however a difficult problem (if not impossible to be solved), taking into account also that the problem of ethics is debated for millennia, and there are several theories on it. For example two of the most important are the deontological (Kant, Kierkegaard and Nietzche) and teleological (Aristotle and followers of utilitarianism) theories. In the former, the ethical dimension of an action is given by the character of the respective action and, in the theological case, only by the result of the action. This distinction is also made in AI programs: a deontological model contains rules, ontologies, or another knowledge base that state what is allowed or not to be done. In the case of the teleological approach, only the ethics of the obtained result is analyzed, the ethics criterion used in actions are not explicit, a situation often encountered in the case of sub-symbolic AI, based on neural networks, unlike the deontological case.

[4] https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence

The Good, the Bad and the Ugly about Explainable and Trustworthy AI

Assistant Professor Alexandru Sorici,
University POLITEHNICA of Bucharest

In recent years we come across increasing talk about *explainable and trustworthy Artificial Intelligence* in academic journals, EU-funded research initiatives and specialized media outlets. The notion is presented as a major stepping stone for the ethical development and use of AI systems.

But what does *explainable or trustworthy AI* really entail?

The EU High-Level Expert Group on AI posits that the lifecycle of an AI system must be based on seven principles. Among those, *explainability* is broken down into: *technical robustness and safety, transparency and accountability*.

Technical robustness and safety means that AI systems need to be accurate, reliable and reproducible. Resilience and failsafe mechanisms have to be built into such systems.

Transparency means that AI systems, as well as the teams that develop them, must be able to explain and communicate the elements involved in their operation: the data, the system itself and its business model.

Accountability is closely linked to the notion of fairness. It requires that a mechanism of *responsibility* be established, both before and after the development and operation of an AI system.

The Good. A recent 2020 report by the Capgemini Research Institute on "*AI and the Ethical Dilemma*" shows that there is increasing awareness of the need and practice of explainability in automated decision-making both among consumers and company executives. Efforts have been made to explain how certain systems work in a language that people can understand. Examples include financial institutions who create "digital alternates" of a customer profile to come up with counterfactual situations (the types of changes in key variables that would have led to a different outcome) when looking at automated loan rejection decisions. Google's "Explainable AI" offering can quantify how each data point contributes to the result[1], while Microsoft's "InterpretML" can show the primary factors that dictate how its machine learning models make a decision[2].

Overall, the share of organizations that make their AI models explainable (at least in the sense of having an *interpretable* narrative) is increasing year by year.

The Bad. The same Capgemini report highlights that, while the general explainability of AI systems is making progress, some of its component areas are lagging behind. Consumer's perception of knowing *what part of their data* was an AI system trained on or whether the system could give *consistent outputs* in

[1] https://www.researchworld.com/can-googles-new-explainable-ai-make-it-easier-to-understand-artificial-intelligence/
[2] https://docs.microsoft.com/en-us/azure/machine-learning/how-to-machine-learning-interpretability

a repeatable manner has decreased in recent years.

Similarly, the report shows that less than half of the organizations they surveyed implement a mechanism that audits the AI systems from an ethics perspective. End-to-end reproducibility (achieving the same output for the same or similar input) is not verified, nor is the case that organizations clearly outline how the AI system was built, what data it was trained on and what data it was tested on, before being put to consumer use.

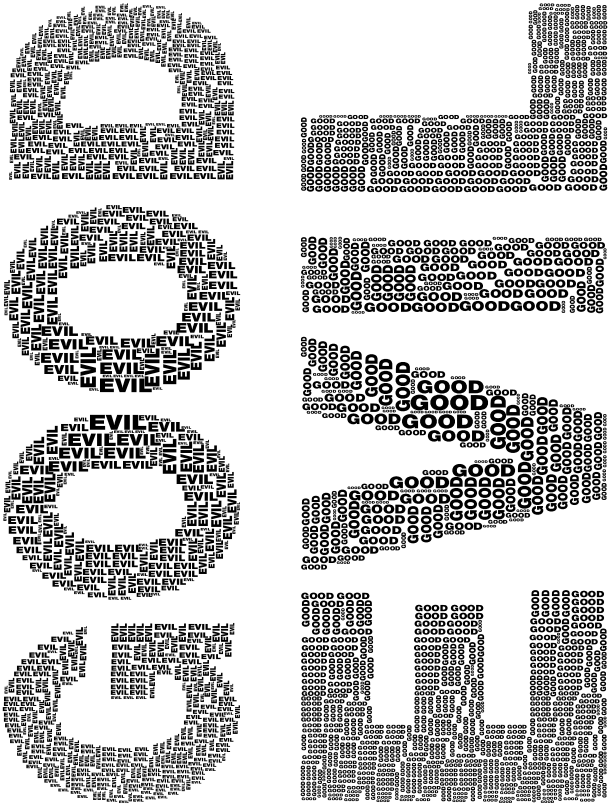
The Ugly. The issues highlighted above can be partially addressed through policy improvement efforts around communicating the process of designing, developing, testing and operating AI systems.

However, an important part of these issues stems from the *growing complexity* of AI models.

While some machine learning models in use today (e.g. decision trees, automated rule extraction models, adaptive explainable neural networks) are amenable to the development of technology tools that allow for transparency into their learned inner workings, the same cannot be stated for complex neural networks, such as the GPT-3 large scale natural language processing model, which have billions of parameters that can influence the final output. In such a context, one can bring transparency into the *inner rules* of the model, but one cannot guarantee that the rules are *interpretable*, that they carry any human understandable meaning.

The Conclusion. The issues outlined previously show why the *explainability of AI systems* is a desired, required yet *complex* issue. Increasing model complexity is the reason why many academics are focusing on building tools for the *interpretability of AI models*, to ensure that people can understand *the impact* of the decisions made by an AI solution.

While the development of such technological tools is an ongoing process, organizations can still increase the robustness, transparency and auditability of their AI systems through avenues taken at a *policy level*. This includes things such as clearly outline the intended purpose of the AI system, embed diversity and inclusion principles throughout its lifecycle, ensure human oversight, as well as putting customers in charge of their AI interactions by empowering them with privacy controls.



Metaverse, artificial intelligence and non-verbal data: ethical interferences

Asist. univ. dr. Mihail-Valentin Cernea,
Bucharest University of Economic Studies

In 1992, Neal Stephenson published the science fiction novel "Snow Crash" in which he paints the vision of a new Internet, in where the digital presence, mediated by screens, keyboards or mice is replaced by the physical and mental presence, an Internet designed as a metaphor for the real world - a metaverse. Almost 30 years later, Mark Zuckerberg announces the reorientation of Facebook's efforts and resources towards creating such a metaverse and changing the company's name to "Meta", bringing the world imagined by Stephenson into reality. It is difficult to imagine that such a virtual space, if successful and, thus, populated by billions of users, can function in the absence of the widespread use of processes involving various applications of Artificial Intelligence. I will briefly explore, in the following article, potential moral issues that may arise at the intersection of the metaverse and AI, at the moment when Big Tech algorithms will have unrestricted access to almost all aspects that make up a human personality. I will end the intervention by introducing a complication determined by the emergence of the so-called brain-computer interfaces.

First, a few details about what exactly such a metaverse is: in principle, we are talking about a vast virtual reality, accessible through specialized headsets that should function as an extension and a possible replacement for the Internet. Instead of websites and web applications, we travel with our own bodies in spaces that perform about the same functions. For example, a visit to a large retailer's website turns into a trip to a virtual mall where we can interact with a 3D simulation of the product we would like to purchase. A meeting on Zoom turns into a conversation between digital avatars in a virtual park. Mark Zuckerberg's dream is not as far-fetched as it might seem, given that the type of VR interactions I described above is already possible with the technology available in the consumer market. Even Meta owns the most popular virtual reality platform on the market, Oculus Quest 2, with around 4.6 million units sold by the first quarter of 2021 alone, according to Counterpoint Research. Among the applications already available on the Quest platform, working in a shared virtual office is a reality.

However, a large part of the public looked at the announcement from Meta with cynicism – it looked like a tactic hoping to change the image of a company accused of a multitude of ethical and legal issues. As various uses of artificial intelligence on the data sets collected from user activity are practically the real business of the former Facebook, it is worth reflecting on the specific ethical dilemmas of AI that become a reality with the emergence of the metaverse.

Given that we are talking about a largely existing technology, modeled on current digital practices, the difference between the moral issues specific to the use of AI in the IT industry in general and the metaverse is only one of degree, not necessarily one of nature.

We can raise issues like those related to social networks or video games:

- respect for users' right to privacy in the context of a much stronger virtual presence;
- issues related to the dependence that such a space can cause once the underlying algorithms determine the preferences of vulnerable consumers
- issues related to algorithmic discrimination and how real-world injustices will be replicated in the virtual world.

Once we realize the size and quality of personal data that can be collected by companies in the metaverse, however, the urgency of a moral debate about the protections that users should enjoy in any future metaverses that will emerge from all corners of Silicon Valley and their effect on social reality is clear. The novelty brought to the data market by the metaverse is a much better ability to collect non-verbal data about users: gestures, facial expressions, body movements, where they focus their gaze, and even deeper data about their feelings. Recent studies on the ability to identify a user based on this type of data show that some algorithms need 5 minutes of training to be able to recognize 95% of individuals about whom they have non-verbal data (Miller et al. 2020). Thus, in metaverse, we will not be able to lose ourselves in the crowd in front of the AI.

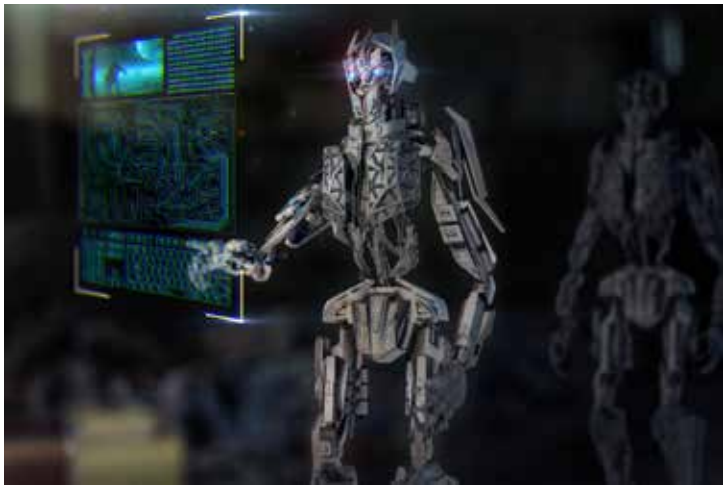
If we add to this complex technological equation the development of brain-computer interfaces that will allow uninterrupted communication between digital environments and the human brain, we can intuit that the use of AI in metaverse will massively increase the power of digital giants to know, but also to control their users.

Photo Credit: PIXABAY

The Ethics of AI Use in Military Applications

Dr. Cristian Ducu
Centre for Advanced Research in Management and Applied Ethics

Photo Credit: PIXABAY



Probably, the first thing that will come up in the mind of most of those that read the title *The Ethics of AI Use in Military Applications* is something similar to the robots we see in the SciFi movies produced in Hollywood. We are far from that moment, but there are a lot of applications and scenarios of technological development that force us to ask ourselves more and more about their ethical implications.

The debate around AI ethics is not always carried on in the academic sphere, where rational arguments seem to be the most important, but also in political, legal and military spheres. The major dilema in this debate surrounds the argument used by some politicians and military experts that there shouldn't be (too many) ethical restrictions in designing and developing military applications based on AI (MinAI). They argue that ethical restrictions would be a considerable disadvantage compared to other countries that invest in this area without having similar moral standards or paying similar attention to ethical arguments.

The most illustrative example for this position comes from Nicolas Chaillan, the first Chief Software Officer of U.S. Department of Defence, who accused in 2021, among other things, that the extensive ethical debates around AI ethics are holding back the United States from investing in AI, similarly to China, and being able to respond to future threats. In the same year, the United Nations asked for the adoption of a moratorium on the use of AI for purposes that might harm human rights. Michelle Bachelet, the U.N. High Commissioner for Human Rights, due to issues related to recognition accuracy, discrimination and protection of privacy, mentioned in particular the facial recognition in real time as a problematic technology.

Such a technology is used by police in many countries and it has been used by U.S. troops in Afghanistan, Iraq and Syria to identify members of terrorist organizations. Other countries, like Israel,

China and the United Kingdom, use similar technology in military operations, too. And, in this particular case of facial recognition, which demonstrated its limitations as well as its benefits, are there sufficient ethical reasons to limit or reconfigure its use in the future in critical places like borders and customs, the scenes of terrorist attacks or conflict zones?

Leaving aside this issue of the weight of ethical arguments compared to those concerning security, there are other military technologies based on AI that have serious ethical and legal implications. For example, there is an extensive discussion at international level on the use of the so-called „autonomous weapons" - military installations that have minimal human coordination and control. Imagine an algorithm-controlled aerial vehicle (UAV) tracking from high altitude a school bus on a dirt road from a region controlled by terrorists. That particular UAV has the capacity to obtain high resolution images and, based on them, to identify, track and engage targets. In this case, when missing a satellite connection with its human operator, the UAV decides that an important target is in the bus — a terrorist leader, for instance — and, consequently, attacks it with a rocket. A first ethical question would concern the responsibility for the decisions made by the algorithms of the UAV: who bears the responsibility, and, implicitly, is accountable for the life and death decision made by the machine? In 2012, many public figures asked the international community to ban these so-called „killer robots". In the situation previously described, that dehumanization of the life and death decision has been pushed back by military officials themselves, who admitted and even insisted on the need for better human governance of lethal decision-making. But, often, the reality in the field is different from imaginary examples or armchair experiments.

A second ethical question is related to the principle of proportionality: is that decision to engage a target a desirable course of action in terms of estimated victims and destruction when compared to the estimated threat? What if, for instance, in the school bus we would have, next to the target we follow, five students? Would that lethal decision still be morally justifiable? Think about a different scenario: a swarm of 12 weaponized drones with active payload identified multiple targets (terrorists) in a building used as a school. The terrorists meet there especially because they count on the fact that the army of a western country would not make decisions that would lead to victims among the students. Only that, this time, the swarm of drones operate autonomously and have to decide to detonate themselves around that building. What we know from the intercepted communication is that the meeting has been requested in order to begin a series of coordinated attacks against civilian targets in several western countries; the threat is imminent. How should that algorithm that coordinates the swarm of drones decide from an ethical standpoint in this context?

These kinds of ethical decisions are difficult to be made by current AI technologies without human input. Most probably, with the evolution of quantum computers, AI will overcome the problem of „dimensionality" (the capacity to treat sets of comprehensive data and to make fast decisions based on „learnt" or „found" information) and, implicitly, will be able to step into a realm of more complex decisions that also involve profound moral aspects. At the same time, this giant leap will lead to new ethical challenges that we are unable to foresee today.

AI, moral externalities, and soft regulation

Mihaela Constantinescu, PhD
Romanian Young Academy,
Research Center in Applied Ethics (CCEA) –
Faculty of Philosophy, University of Bucharest

As deployment of Artificial Intelligence (AI) that relies on machine (deep) learning faces more and more ethical challenges, the need to approach AI from a robust ethical framework is more and more pressing. This has led international bodies such as OECD, the European Commission or UNESCO, to develop policy documents that integrate and respond to ethical concerns through proposed strategies, tools, and mechanisms. But despite the positive outcome envisaged, many ethicists are still reluctant to the way the ethics guidelines will be used in practice, wondering whether all this will not simply amount to ethics washing (Floridi, 2019).

As some highlight, mere existence of ethical guidelines will not generate the expected outcome in the industry, as their recommendations will simply not be fulfilled in lack of monitoring (Hagendorf, 2020).

So how can we make sure that policy meets practice, namely, that ethical guidelines for AI deployment will indeed be respected across markets? How should we best mitigate harmful effects of AI use? While there is no straightforward answer to such questions, I would like to address part of it by connecting the deployment of AI to the concept of moral externalities.

Photo Credit: PIXABAY

AI moral externalities. In what follows, to discuss the impact of AI deployment and relevance of ethical guidelines for practice, I refer to the concept of *AI moral externalities* and highlight issues related in particular to negative moral externalities of AI. Briefly put, by AI moral externalities I understand the collateral moral damages and benefits of deploying AI, which are borne by a third party that has no responsibility for either deployment or use of AI. This understanding builds on the definition used in the field of business ethics, where negative moral externalities refer to "morally significant consequences that seem to escape ethical reckoning about what is owed by an actor – situations that defy our capacity to assign responsibility for preventable harm" (Gowri, 2004: 40).

It is quite clear that the concept of moral externalities differs from its initial use in economics - moral externalities are not a subset, but rather a parallel phenomenon to economic externalities, because, for instance, there is no necessary financial equivalent to the moral costs or benefits (Gowri, 2004). As a result, economic treatment of externalities meant to internalize the cost of externalities through taxes or hard regulation, does not have a direct equivalent strategy when it comes to moral externalities. Therefore, discussing the harmful collateral effects of AI deployment in terms of moral externalities, and not economic externalities, might help us better realize the fact that AI externalities cannot simply be equated with some financial costs and straightforwardly be internalized by emitters through, for instance, hard regulation. So where does this leave us in terms of dealing with AI moral externalities and ensuring that ethical guidelines for AI deployment will indeed be followed across markets?

Regulating AI deployment. Despite on-going discussions regarding best means to regulate AI deployment (Reed, 2018; Taddeo and Floridi, 2018), there is no consensus over how this can be practically achieved. First, things are complicated by the very fact that it is often not very clear what it is to

be regulated (Almeida et al., 2020). Second, legislating morality is not a simple task, and it remains one of the important issues in moral, political, and legal philosophy (Hatzis, 2015). Third, regulation is limited in its ability to capture proactively the broad set of harmful outcomes resulting from fast-moving industries such as robotics and Artificial Intelligence.

Nonetheless, letting the industry regulate itself without a common ethical framework and without oversight might result in unforeseeable harmful effects on individuals, who often do not have the information, tools, or power to exert a real market pressure. Current and in-progress AI ethical guidelines might indeed provide a common ethical framework. However, they risk being unnecessary without proper operationalization in measurable results, providing a surveillance tool for regulators and even the general public.

AI, soft externalities & soft regulation. AI moral externalities are a form of soft externalities (to use the term put forward by Epstein (1997) when speaking of social externalities) and may thus require soft regulation, such as ethical guidelines. However, as discussed above, this may not be enough. A possible solution is to add supplementary soft governance tools, such as specific (open) ethical standards (e.g., the IEEE-P7000 suite and certification programs; see IEEE, 2019). Given their practical application (Winfield et al., 2021), such standards may operationalize AI ethical guidelines and make them measurable, with the advantage that, if necessary, standards may also be enforced through hard regulation in the future (Theodorou & Dignum, 2020).

To properly mitigate harmful effects of AI deployment and use, the combined soft regulation

BRIEF OVERVIEW OF AI MORAL EXTERNALITIES

Level of AI deployment	Ethical issues	AI moral externalities
AI production	*bias* embedded in data sets used for training AI algorithms	polarization, radicalization
	global data gathering, mining, extraction and use of *big data*	private data exposure, digital reconfiguration of knowledge and truth
	use of natural resources & low-cost labour markets for AI hardware production	exploitation of human labour, environmental damages
The nature of the AI itself	the moral status of highly autonomous AI systems	decrease in human control, autonomy, and responsibility
Applications and uses of AI	political authority & control, individual freedom	negative impact on minorities – bias, discrimination
	prediction algorithms (insurance, health, surveillance etc.)	denied access to medical insurance, reinforced bias in recruitment
Impact for non-owners and non-users of AI	local, regional, national and global use of AI (especially in public administration, health)	non-user / non-owner discrimination, increased inequality, economic & social divide, inequitable access

tools of ethical guidelines and ethical standards need to consider a broad range of AI moral externalities. The table briefly sketches the realm of AI moral externalities along the following phases of AI deployment: (1) AI production (2) the nature of the AI itself (stand-alone or part of a robotic AI system) (3) applications and uses of AI (4) impact for non-owners and non-users of AI.

Finally, future efforts towards actual alignment between soft regulation and AI deployment will need to consider ways to correlate AI moral externalities with the complex, intertwined network of moral responsibility of all those involved along the entire AI deployment cycle, with special attention paid to collateral moral harms generated for those who are outside this network.

Photo Credit: PIXABAY

Bibliography

Almeida, P., Santos, C.D., & Farias, J.S. (2020). Artificial Intelligence Regulation: A Meta-Framework for Formulation and Governance. *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 5257-5266.

Epstein, R. A. (1997). Externalities Everywhere: Morals and the Police Power. *Harvard Journal of Law and Public Policy*, 21, 61-69.

Floridi, L. (2019). Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. *Philosophy & Technology*, 32, 185–193.

Gowri, A. (2004). When Responsibility Can't Do It. *Journal of Business Ethics*, 54, 33-50.

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Mind and Machines*, 30, 99-120.

Hatzis, A. N. (2015). Moral Externalities: An Economic Approach to the Legal Enforcement of Morality. In Aristides N. Hatzis & Nicholas Mercuro (Eds.), *Law and Economics: Philosophical Issues and Fundamental Questions* (pp. 226-244). London/New York: Routledge.

IEEE (2019). *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems*. First Edition. Piscataway, NJ: IEEE Standards Association. Tech. rep.

Reed, C. (2018). How should we regulate artificial intelligence? *Philosophy Transactions of the Royal Society*, 376.

Winfield AFT et al. (2021). IEEE P7001: A Proposed Standard on Transparency. *Frontiers in Robotics and Artificial Intelligence*, doi: 10.3389/frobt.2021.665729.

Taddeo, R. & Floridi, L. (2018). How AI can be a force for good: An ethical framework will help to harness the potential of AI while keeping humans in control. *Science Review*, 361, 751-752.

Theodorou, A. & Dignum V. (2020). Towards ethical and socio-legal governance in AI. *Nature Machine Intelligence*, 2, 10-12.


AI