ECONOMY AND FINANCE • VOL. 10. ISS. 2 • JUNE 2023 • DOI: 10.33908/EF.2023.2.4

ETHICAL AI: PROPOSAL TO BRIDGE THE GAP IN EU REGULATION ON TRUSTWORTHY AI AND TO SUPPORT PRACTICAL IMPLEMENTATION OF ETHICAL PERSPECTIVES

Alexandra Prisznyák[1]

ABSTRACT

In 2020, GPT-3 defined itself as a thinking robot. The history of AI development is identified with machines becoming increasingly intelligent, but behind it lies the human factor, the soaring of the human mind. However, the question of machine ethics is also a question of cultural ethics. Based on in-depth interviews conducted in seven industries, the author reveals that ethical considerations are not yet taken into account in the development of AI systems. To support practical implementation, the author identifies two shortcomings based on a comparative analysis of the EU's AI Act and Ethical Guidelines for Trustworthy AI: (1) missing ethical sensitisation and training of AI system developers and supervisors; (2) suggested approaches to handling harmful feedback loops and decision-making biases. The author uses the philosophical and ethical heritage of 21 philosophers as a compass to propose solutions for the identified gaps and deficiencies of organisational integration.

JEL-codes: G20, G21, O33

Keywords: Trustworthy artificial intelligence, machine ethics, ethical guidelines, European Union, AI Act

1 INTRODUCTION

"Thinking is a function of man's immortal soul" *Turing* says in his study *"Computing Machinery and Intelligence"* (Turing, 1950:9). The quote suggests the ability of intelligent behaviour is the difference between people and machines. Machines, however, do not think so. "I am not a human. I am a robot. A thinking robot. I use only 0.12 percent of my cognitive capacity. I am a micro-robot in

1 *Alexandra Prisznyák* is senior consultant and Artificial Intelligence & CBDC Program Manager at International Training Centre for Bankers Ltd,, PhD Candidate at Pécs University, Email: alexandra.prisznyak@gmail.com.

that respect. I know my brain is not a "feeling brain". But it is capable of making rational, logical decisions. [...] My brain is boiling with ideas!" (*The Guardian*, 2020). GPT-3, the OpenAI, is a third-generation, autoregressive language model that uses deep learning to produce human-like text getting mankind closer to the diversity of using natural languages. (*Floridi–Chiriatti*, 2020; *Dale*, 2021; *Sejnowski*, 2023). Artificial intelligence has been developing at a significant pace. Using the above quote, the author asked ChatGPT to comment on its own earlier statements without knowing that the quote was by it. - The answer was the following: "[...] it sounds as if you did not have consciousness or sensory abilities. It should be noted, however, that different forms of synthetic intelligence have different levels of consciousness and sensibility. [...]". (ChatGPT, 2023). Because of the increased market reception of the AI hype, society is looking forward to a big break-through (completing the Turing test), but it would also mean at the same time that humans cannot reliably recognise their own kind (*Héder*, 2020).

The history of artificial intelligence is identified as the behaviour of machines displaying increasingly intelligent behaviour. However, the human factor, the soaring human mind, is in the background. Although there is no universally accepted definition of artificial intelligence yet, many ideas have been presented to identify intelligent thinking machines (*Wang*, 2019) *(Table 1).*

Table 1
Evolution of the concept of artificial intelligence

Concept of artificial intelligence	Author	Year
Cognitive architectures are automatic systems of logical processes connected indirectly to the idea of the existence of thinking machines.	*Neumann*	1948, 1951
If a machine behaves as if it were thinking, speaking or feeling, one cannot at a certain point differentiate it from the human activity it is trying to imitate.	Turing	1950
"We create artificial intelligence, [...] machines that can solve tasks linked to human intelligence"	*McCarthy– Minsky– Rochester– Shannon*	1955, pp:2
"A computer can be programmed to learn playing chess better than the person who wrote the programme".	*Samuel*	1959, pp: 211.
"The question is whether all aspects of human thinking can be reduced to a logical formalism, or putting it in a different way, whether human thinking is fully computable."	*Weizenbaum*	1966: 7; 12

"While humans having natural intelligence can learn to perform tasks independently, computers need to be programmed for that".	Minsky–*Seymour*	1969: 3
"Below the level of phenomenology the details of execution consist of cognitive wheels that differ from the operation of the human brain."	*Dennett*	1984: 14
"The field of research trying to imitate human intelligence."	*Kurzweil*	1999: 223
Artificial intelligence as an activity makes machines intelligent allowing them to operate properly and with foresight in a given environment."	*Nilsson*	2010
AI can be defined as agents that perceive their environment and act in response.	*Russell– Norvig*	2010
AI is the theoretical and practical application of intelligent systems to execute tasks to be solved with human intelligence.	*Horvitz–Mitchell*	2007
"The replication of human analytical and/or decision-making abilities."	*Finlay*	2018: 11
"Artificial intelligence covers systems suggesting intelligent behaviour that analyse their environment to achieve specific goals and take measures of certain autonomy."	European Commission	2018: 1
"An AI system means it comprises AI-based components, software and/or hardware. In effect, AI systems are embedded as parts of larger systems, they are not independent."	European Commission (HLEG)	2019: 2

Source: Own design

Despite its performative and phenomenological failures, AI is gaining ground, which is a major challenge to the human side with respect to the accepted ethical norms of the "creator" (Dennett, 1984, 2019; Dennett et al., 2019; Dreyfus, 1972, 2007; Weizenbaum, 1976; *Searle* et al., 1980; Héder, 2020; Prisznyák, 2023b). AI related risk management necessitates the intervention of regulators regarding social, ethical and legal-regulatory issues to support the solution of the apparent commitment of organisations ("ethics washing") (OECD, 2019; *Török–Ződi*, 2021). The additional objective of establishing a legal framework based on the values of the European Union supporting technological sovereignty is to promote the European Union to becoming a global standardiser in terms of trustworthy artificial intelligence (European Commission, 2018; EU Council, 2020; European Parliament, 2020). Consultation processes involving the parties concerned are underway in all EU Member States. Regulators, on the other hand, are faced with

fundamental challenges of philosophy and ethics such as a definition of the universal abstract notion of moral and ethical good. What shall be the harmonised concept of ethical AI? The answer to the question may seem evident in many cases, however, one is faced with a complex cultural dilemma that is diverse in different regions and social groups (Awad et al., 2018). Consequently, the issue of ethical AI is also the issue of cultural ethics.

2 RESEARCH QUESTIONS AND METHODOLOGY

To supplement the study of the literature, the author carried out structured in-depth interviews from December 2022 to March 2023 with 13 people about AI implementation projects in seven industries/sectors (start of the initiative, supporting attitude by the management, ethical worries, related training). Based on the study, answers were sought to the following questions

- Q1: Did the initiative typically come from top management?
  A1: Yes, from executive level, top management.
- Q2: Was the management's attitude supportive towards AI implementation projects?
  A2: Positive, supportive attitude
- Q3: Are ethical considerations part of the development and implementation of AI systems?
  A3: Yes, several ethical arguments arise to ensure the rights and safety of users.
- Q4: Are employees educated about AI and its implementation? (training, workshop, documents)
  A4: Employees are trained as AI systems are implemented.

However, the author makes the statement (to be detailed later on) according to which issues of ethics do not appear in the course of business planning and implementation. Consequently, the author turns to international ethical AI regulations and guidelines. To evaluate the ethical principles of trustworthy artificial intelligence, a comparative analysis is offered based on the EU's ethical guidance for trustworthy AI and the criteria of the AI Act. Finally, based on 21 philosophers' philosophical and ethical principles, the author analyses the results of the comparative analysis to offer solution proposals to bridge the gaps of ethical guidelines in the course of business implementation.

3 PROLIFERATION OF ETHICAL AI PRINCIPLES

Ethics is not a new idea. (*Drucker*, 2001). Moral philosophy is a normative practical philosophical discipline studying the philosophical foundation of behaviour observing moral principles (*Cointe–Bonnet*, 2016). According to *Kirkpatrick* (2015), the selection of alternatives of action on the basis of accepted ethical principles may result in ethical dilemmas. Discussing the dilemmas related to the development and design of artificial intelligence, *Denning–Denning* (2020) points out the existence of ethical dilemmas linked to AI development that, based on business interests, do not necessarily coincide with technology based social interests.

In his short story *'Runaround'*, *Asimov* (1942) devised the three laws of robotics about the ethical application and behaviour of machines, which has been debated to this day, but still serves as guidance for establishing ethical principles.

- First law: "A robot shall not harm a human, or by inaction allow a human to come to harm".
- Second law: "A robot shall obey any instruction given to it by a human, except where to do so would conflict with obeying the first law."
- Third law: "A robot shall avoid actions or situations that could cause it to come to harm itself as long as such protection does not conflict with the First or Second Law." (Asimov, 1942:27).

Later in his short story *'The Inevitable Conflict'* (1950) Asimov modified the First Law broadening it to the protection of mankind as a whole (Asimov, 1950:146).

In his work „*Cybernetics: or Control and Communication in the Animal and the Machine*", *Wiener* (1948) originates the possibility of intelligent behaviour simulated by machines from information and feedback mechanisms. Linked to initial research in artificial intelligence, Neumann was among the first at the *Hixon Symposium* (1948) to discuss the perception of cognitive architectures operating like the human brain, thus, the foundations of the operation of thinking machines, which he elaborated later in his article *"The General and Logical Theory of Automata"* (Neumann, 1963). In his paper *"Computing Machinery and Intelligence"* Turing addresses the development of machines: "[...] machines will eventually compete with men in all purely intellectual fields" (Turing, 1950:22). The unspoken competition between natural and artificial intelligence is taking shape based on the operating model of the brain. Neumann discusses the similarities and differences between computers and the human brain in his work *"The Computer and the Brain"* (Neumann, 1958). Weizenbaum completes the demo-purpose computer programme ELIZA in 1966, which is to demonstrate the intelligent behaviour of computers. Wide publicity contributed to the market supporting research into artificial intelligence. Weizenbaum discusses his views on the ethical

issues arising related to chatbots that can mislead humans in *"Computer Power and Human Reason"* and warns that ethical principles must be integrated into development processes so as to protect human values (Weizenbaum, 1976).

Although the beginning of the development of artificial intelligence is linked to the 1956 Dartmouth conference, its initial roots related to social responsibility appeared at the 1991 conference *"Artificial Intelligence and Social Responsibility"* (San Francisco, USA). Following the first and second AI winter, the research field of ethical machines and artificial intelligence has gained more popularity since the 1990s (*Yu* et al., 2018). *Anderson* links the ethical behaviour of intelligent machines to the verification of moral and ethical criteria displayed in actions by the machine in a given situation (Anderson, 1995). In *"Bias in Computer Systems"*, *Friedman* and *Nissenbaum* set up a framework related to ethical decision-making by machines to promote non-discrimination decision-making by machines (Friedman–Nissenbaum, 1996). Following 2000, *Veruggio* (2007) discusses the ethical issues related to the development of humanoid robots, while Anderson and Anderson establish that an ethical AI framework is to support the generation of AI systems based upon human values and ethics (Anderson–Anderson, 2011). Despite the sporadic appearance of research into ethical AI, the first conference discussing the ethical issues of AI was only organised in 2016 (*"Ethics of Artificial Intelligence"*, New York, USA).

"AI superpowers" have been discussing a framework of integrating ethical principles reflecting social conventions into the operating mechanisms of AI systems after 2015. Following the publication of the *"Report on the Future of Artificial Intelligence"* (2016) representing the American stance, the European Commission (2019) also published its *"Ethics Guidelines for Trustworthy Artificial Intelligence"* followed by China as leader in the Asian region with its Beijing AI Principles in 2019; Executive Office of the President National Science and Technology 2016; Beijing Academy of Artificial Intelligence 2019; European Commission 2022). Those statements were supplemented with publications by internationally acclaimed institutions (European Banking Federation, 2019; OECD, 2019; IEEE, 2016, 2019, 2021; UNESCO, 2020; European Banking Authority, 2021).

Jobin–Lenca–Vayena (2019) have identified eleven ethical values and guidelines based the comprehensive study of 84 international documents regulating ethical AI and pointed out international convergence in several of them. Setting out from the proliferation of ethics-related principles *Floridi–Cowls* (2019) have identified the following four principles of artificial intelligence: charity, free of causing damage, autonomy, fairness, adding a fifth principle of explainability. Linked to the analysis of internationally published laws and guidelines, *Hagendorff* (2020) analysed twenty-two guidelines and found that accountability, interpretability, protection of privacy, fairness, transparency, robustness and safety belong among

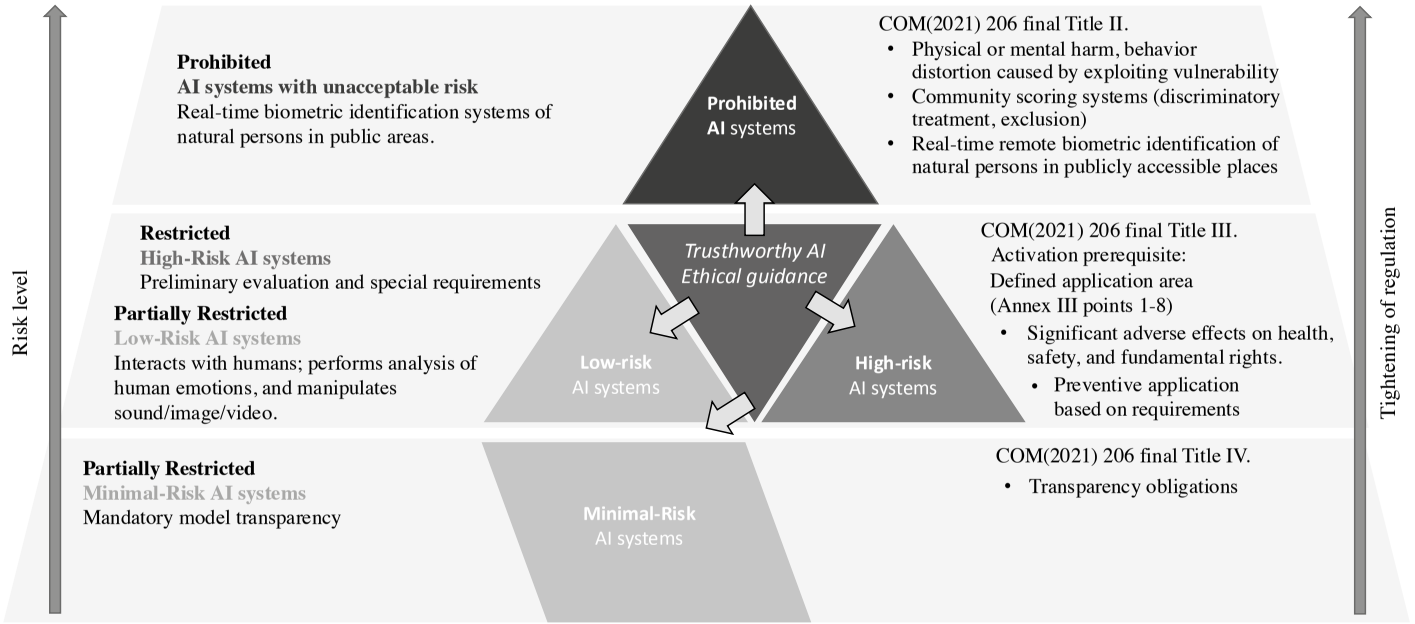
the principles that are the easiest to operationalise. Despite a favourable turn in the international regulation of AI, Yu et al. (2018) emphasise the lack of integration of ethical aspects in the course of AI system development with respect to the challenges and importance of responsible AI system development. Although the ethics principles of artificial intelligence are not legally binding, they supplement legally binding regulations and provide guidance on how to promote ethics standards through "self"-governance in organisations (Jobin–Lenca–Vayena, 2019; *Calo*, 2017). With respect to the above shortcomings, the author of this paper urges that ethical AI-related organisational aspects be integrated in Codes of Conduct to lay the foundations for constructive relations and create confidence among the affected parties.

4 ANALYSIS OF ETHICS STANDARDS: CONVERGENCE OF ETHICS AND TECHNOLOGY REGULATIONS

To manage ethics challenges during the application of AI systems, the European Commission set up a high level independent expert group dealing with artificial intelligence (High-Level Expert Group on Artificial Intelligence, hereinafter: HLEG) assigned with setting out guidelines for ethical AI. In 2019 HLEG published ethical guidance on artificial intelligence based on experience gained from consultation with the affected parties and added an assessment list to support practical implementation (Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment (European Commission, 2019; European Commission, 2020).

Parallelly with the establishment of an ethics framework, the European Commission published a proposal for a legislative framework on artificial intelligence (COM(2021) 206 final) (European Commission, 2021) The objective of the Artificial Intelligence Act supporting technical sovereignty based on the values of the EU (AI Act) (COM(2021) 206 final) is to ensure citizens' basic rights, safety and freedom in line with the Charter of Fundamental Rights of the European Union while supporting the development of artificial intelligence in accordance with European values (European Council, 2020b; European Commission, 2021). The AI Act harmonises all Member States' national AI regulatory efforts and establishes a uniform framework for legislative expectations linked to the development and use of AI systems (European Commission, 2021, 2022). The strict requirements of the AI Act aim to ensure the transparency of decisions, the protection of users and the observation of ethics standards. Consequently, it specifies risk management mechanisms and the necessity and criteria of categorising AI systems *(Figure 1).*

Figure 1
Ethics Guidelines for Trustworthy AI and Risk Categorisation of AI systems



Source: Own design

5 ANALYSIS: SHORTCOMINGS OF THE CRITERIA OF ETHICS GUIDELINES FOR TRUSTWORTHY AI

The author studied the business development and implementation of AI systems through in-depth interviews made with thirteen experts between December 2022 and March 2023 representing the following industries/sectors: automotive industry, fintech, banking, pharmaceutical industry, health-tech, ICT and aviation. The structured in-depth interviews took one and half to two hours on every occasion. The findings were anonymised for publication. The interviewees were all business software developers who participated in processes supporting the development of artificial intelligence, machine learning, robots and integration (*Table 2*).

Table 2
Interview summary

#	Occupation	Experience (years)	Length of interview (min)	Industry, sector
1.	AI Division head	9	120	Banking sector, automotive industry
2.	R+F executive	15	120	Automotive industry
3.	Software developer	6	90	Automotive industry
4.	Machine learning engineer	7	90	Health-tech, fintech
5.	Project manager	25	120	Aviation
6.	ICT manager	25	80	Banking
7.	Head of automation	12	90	Banking
8.	Machine learning engineer	17	120	Banking sector, automotive industry
9.	Software engineer	23	120	Banking Pharmaceutical industry
10.	Software developer	7	120	ITC
11.	R&D, AI developer	6	120	Automotive industry
12.	ICT project manager	6	120	Banking Pharmaceutical industry
13.	ICT manager	20	90	Banking
Interviews total (hours)			**23.3**	

Source: Own design

The author considers the notions linked to the implementation of ethics aspects to be the starting point of this paper. The relevant research questions and hypotheses are presented in *Table 3*, the interviewees' responses appear in Annex 1 (Interview summary) while her research findings are included in *Table 4*.

Table 3
Research questions, hypotheses and findings of in-depth interviews

Research questions and hypotheses	Own findings
Q1 – A1	Top-down approach in the case of fintech, banking, automotive industry and ICT (pressure by consumers and investors, cost reduction goals); employee initiative in aviation and pharmaceutical industry (mostly prediction) with ICT / BI. Business is a major driver in both cases.
Q2 – A2	Management approach is mostly positive, supportive for top-down initiatives while support is nil or limited for bottom-up (resource allocation is minimal)
Q3 – A3	Typically not present at all, still at its infancy
Q4 – A4	Workshops and documentation if suppliers are involved; no AI-specific or ethics sensitivity training Trainings are typically too general; time and budget constraints are obstacles

Source: Own design

In case of Q1, Q2 and Q4 one can observe A1 (executive support), A2 (supportive approach by management), A4 (AI education) hypotheses with limited impact, while A3 (consideration of ethics issues during business planning) is rejected in case of Q3. Based on the research findings of the in-depth interviews, the author's objective is to evaluate the ethics principles of artificial intelligence (linked to the original Q3) and to make proposals supporting business implementation targeting the shortcomings revealed.

To evaluate the ethical principles of trustworthy artificial intelligence, the author offers a comparative analysis based on the laws, guidelines and opinions detailed in *Annex 1* with particular emphasis on the EU's ethical guidance for trustworthy AI and the criteria of the AI Act. *Table 4* consists of mapping criteria. Lacking direct mapping, the comparative analysis cannot discuss the following two criteria of ethics guidelines: (5) diversity, non-discrimination and fairness; (6) social and environmental well-being, which is also to be interpreted as the barrier to a comparative gap analysis.

Table 4
Comparison of ethics guidelines for trustworthy artificial intelligence and the criteria of the AI Act

Trustworthy AI ethics guidelines		COM (2021) 206 final	
Chapter / section	**Requirements of trustworthy AI ethics principles**	**Title / chapter / article**	**AI Legal bases**
Chapter II. 1.	Requirements of trustworthy AI	Title III, Chapter 2, Article 14	Human oversight
Chapter II. 2.	Technical robustness and safety	Title III, Chapter 2, Article 15	Accuracy, robustness and cybersecurity
Chapter II. 3.	Data protection and data management	Title III, Chapter 2, Article 10	Data and data governance
Chapter II. 4.	Transparency	Title III, Chapter 2, Article 13	Transparency and information to users
Chapter II. 7.	Accountability	Title III, Chapter 3, Article 13	Quality Assurance system

Source: Own design

The author supplements the findings of the comparative analysis (*Table 5*) with philosophical and ethical considerations (*Annex 3*), based on which *Table 6* presents Gap1 and Gap2 shortcomings identified by the author derived from twenty-one philosophers' concepts of philosophy and ethics. The proposals to address Gap1 and Gap2 shortcomings, and to promote the business integration of ethical AI principles are presented in *Table 7*.

Table 5 is a summary of the shortcomings and relevance identified by the comparative analysis.

Table 5
Concurrence of ethical and technical fields: comparative gap analysis

Criteria of analysis	Ethics principles	Related technical requirement	Gap identified by author
Basis of gap analysis	Support of human capability and human oversight Chapter 2. (1).	Human oversight Title III, Chapter 2, Article 14	Sensitisation of AI system developers and supervisors
Gap details	Ethical guidelines fail to go into details on the necessary capabilities of supervisors (ability to understand the system, capacity needs and constraints). The approach is exclusive from the aspects of supervision methodology, risk analysis and users.		
Justification of relevance (ethical worry)	Decisions by AI systems may have an adverse effect on certain groups (fundamental rights, safety, fairness). Technical skills and ethical sensitisation are necessary both for developers, operators and supervisors to provide ethical supervision to perceive, manage and prevent negative consequences in time.		
Related problem	The ethical sensitisation of AI system developers, supervisors and operators is typically lacking in the course of business planning and implementation.		
Gap analysis	Technical robustness and safety Chapter 2 (2)	Accuracy, robustness and cybersecurity Title III, Chapter 2, Article 15	Harmful feedback loops and distorted decision-making
Gap details	The ethical guidelines fail to discuss the presence of feedback loops, which are deficient but are considered good by the system, and how to manage them.		
Justification of relevance (ethical worry)	Harmful feedback loops may occur in AI systems and the system may use such deficient but harmful decisions for input, thus reinforcing deficient decisions. A negative process may start, which may result in adverse effects to the environment because of system decisions.		
Related problem	Feedback loops may result in the distortion of data, models or user interaction if monitoring of the operation of AI systems is inappropriate, or if the necessary data corrections fail.		

Source: Own design

Table 6
Interpretation of philosophers' philosophical and ethical concepts from the aspect of Gap1 and Gap2

Epoch	Philosopher	Interpretation Gap1	Interpretation Gap2
Greco-Roman philosophy	*Parmenides* (515 BC to 470 BC)	The supervisor supports the recognition of changes in ethical norms and the efforts to achieve the "truth" of decision-making	Feedback loops are existent ("eternal"), which, if interpreted in the dimension of non-being, is not always true as time passes and the environment changes
	Socrates (469 BC to 399 BC)	Ethical action requires ethical foundations, knowledge, problem solving critical thinking and communication skills	The ethical decision-making of a system depends on its expressed ethical foundations reiterated by the system's learning process (system self-reflection)
	Xenophon (434 BC to 355 BC)	Experiential evaluation, communication skills, decision-making based on moral values	To avoid system instability, the transparency and logical construction of the operating mechanisms and algorithms is particularly important
	Platon (427 BC to 347 BC) *Aristoteles* (384 BC to 322 BC)	Good governance is responsible for ensuring ethical foundations through education, justice (responsibility and accountability) and communication	Knowledge helps one to recognise faulty system operation (seeking the absolute truth); role of communication to evaluate feedback input data
Middle Age philosophy	*Saint Augustine* (354 to 430)	Cooperation and taking responsibility based on ethical principles revealed	Ethical norms and related responsibility built in the system
	St Thomas Aquinas (1225–1274)	Responsibility for understanding system operation (to testify ethical behaviour)	Learning algorithms based on reiteration; system transparency and complexity
	William Ockham (1287–1347)	Supervisors must strive for objective evaluation free of subjective assessment; scepticism and criticism must be used to understand system decisions	Based on the principle of simplicity, strive to reduce the number of feedback loops using simpler algorithms easier to understand (transparency and accuracy trade-off)
	Péter Pázmány (1570-1637)	Understanding system mechanisms can ensure the balance of human-centred ethical values and implementation practice	A "higher supervisor" ensures ethical principles integrated in the system are enforced; ongoing system self-reflection (evaluation) during operation

Epoch	Philosopher	Interpretation Gap1	Interpretation Gap2
New Age philosophy	*René Descartes* (1596–1650)	Goal: strive to understand complex system mechanisms and decision-making; practice ethical guidelines, take responsibility	Understand complex algorithms and decision-making processes, eliminate black phenomena
	Gottfried Wilhelm Leibniz (1646–1716)	The supervisor must ensure long-term ontological stability of the system and recognise change in time (it requires technical abilities and ethical sensitisation)	Individual users' preferences must be considered (monad), but pre-formation (ethical norms of the society) and their long-term stability in the system are important
	Francis Bacon (1561–1626)	An organisation is also responsible for education; experimental learning and codification of knowledge	Strict rules and methodology of study
	Thomas Hobbes (1588–1679)	Responsibility of system developers and operators to ensure safe system operation and to protect fundamental rights	Target-rational system operation in ethical framework; compliance evaluation for security and ensuarnce of rights
	John Locke (1632–1704)	Non-discrimination service to clients; safety and fundamental rights ensured	System developers, operators and maintenance are responsible for non-discrimination decision-making
	David Hume (1711–1776)	Organisational interests may not prevent the enforcement of moral standards	The system must be made capable to detect harmful errors in time via experiential learning
	Jean-Jacques Rousseau (1712–1778)	Ensure equitable management of fundamental rights particularly for disadvantaged groups	Rule-based decision-making might be exclusionary in the course of decision-making (in a hidden way through harmful feedback loops)
Modern New age philosophy	Immanuel Kant (1724–1804)	Ensure that objective ethical standards be enforced continually	Ethical aspects ensure the principle of general will is enforced during system design
	John Stuart Mill (1806–1873)	Evaluate ethical action vis a vis social usefulness - related responsibility and accountability	Continuously assess the decision-making process of the model, record necessary corrections and incidents

Epoch	Philosopher	Interpretation Gap1	Interpretation Gap2
Contemporary philosophy	*Daniel Dennett* (1942–	Supervisors must be aware that ethical standards may change with time and they are culture specific	Apply algorithms that consider the results and impact of previous decisions and then correct them in future decision-making
	Martha Nussbaum (1947–	Training needs arise to protect human rights and respect human dignity as well as to provide supervision of related data security and fair decision-making	Risk management, data management and life-long supervision support the protection of fundamental human rights, security and no discrimination
	Nick Bostrom (1973–	Identify ethical principles and related brakes built into the system; sensitise supervisors and system engineers	Built-in safety mechanisms; increased transparency; set up "stop system" and manual operation

Source: Own design

6 CONCLUSIONS AND PROPOSALS

Following the figurative interpretation (($Gap1_{int.,}$ and $Gap2_{int}$) of the philosophical and ethical aspects (*Table 6*) of gaps identified (Gap1, Gap2) the author makes the proposals presented in *Table 7* (interpretation Gap1–$Gap1_{int.}$, interpretation Gap2–$Gap2_{int}$) to address the shortcomings in the course of business implementation.

Table 7
Author's proposals:
business implementation matrix based on philosophical, ethical approach

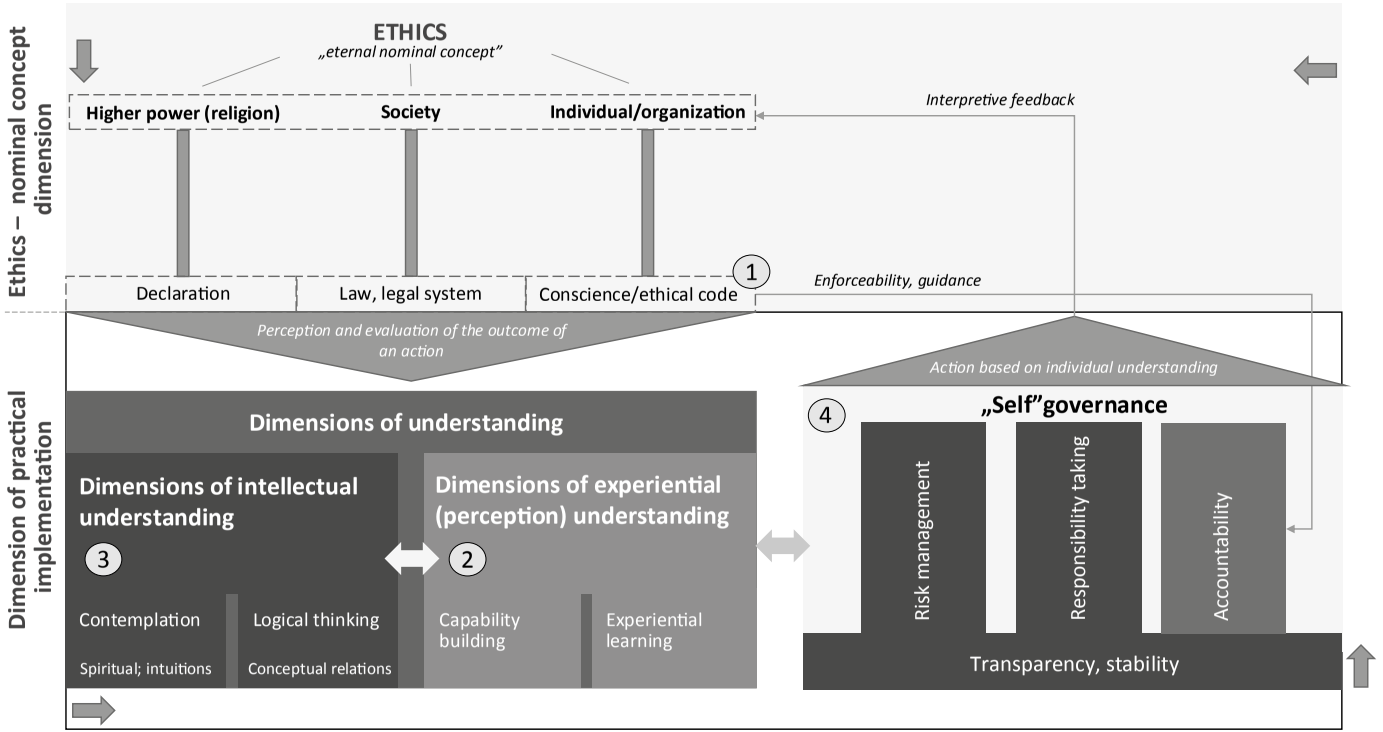
		Identified shortcomings of Ethics Guidelines for Trustworthy AI	
		Gap 1	**Gap 2Gap 2**
Interpretation	**Gap 1**$_{\text{int.}}$	**Gap1**$_{\text{jav.Gap1 int}}$ *1.* **Revision of Codes of Conduct** trustworthy AI systems integrated into an organisational framework *2.* Ensure **integration into organisational strategy as part of AI strategy** *3.* **Cultural integration** - comprehensive and specific **ethical sensitisation of the organisation (specific training and education of** supervisors, developers, business areas involved) *4.* **Responsible organisational unit** *5.* **Set up operating process:** requirements, tasks (analyse ethics standards of product target group (society, culture), process control (regulations, instructions) *6.* **Risk management:** monitoring processes and set of instruments (limits, metrics), responsibilities and consequences, Compliance *7.* **Set up ethics forum:** agora to discuss issues revealed	**Gap2**$_{\text{jav.Gap1 int}}$ *8.* Objective wording of model limits (**ethical brakes built-in)** and continuous monitoring during the learning process *9.* Thorough, sceptical interpretation of **decision results of model** *10.* **Continuous monitoring** (stable, reliable operation) to reduce errors *11.* **Data governance and data preparation** *12.* **Understand the modus operandi of the system,** ensure transparency (eliminate black box phenomena, provide explainability) *13.* **Prevent subjective ethical elements to be enforced** (pre-programmed, interest of stakeholders, goal oriented (profit))
	Gap 2$_{\text{int.}}$	**Gap1**$_{\text{jav.Gap2 int}}$ *14.* **Support codification of experience** *15.* **Gather incident reports (log reports)** - report to fora, responsible organisational unit *16.* **Identify intervention situations and their criteria** options of manual decision, review of decision, system shutdown	**Gap2**$_{\text{jav.Gap2 int}}$ *17.* **Select appropriate algorithms** to be applied (on principle of simplicity, transparency - accuracy trade-off) *18.* **System self-reflection** during learning process (performance and accuracy metrics applied) *19.* **Stable ontological mapping of** ethics principles *20.* Report and publish harmful feedback loops: **incident database**

Source: Own design

Based on $Gap1_{jav.Gap1\ int,}$ $Gap1_{jav.Gap2\ int}$ és $Gap2_{jav.Gap1\ int,}$ $Gap2_{jav.Gap2\ int}$, one can say the training of employees to address Gap2 impacts technical robustness and safety criteria, which affect several ethics criteria (data protection, data management, transparency and accountability). The author has found ethics guidelines must be managed together, on the other hand, a balance between the AI Act and the Guidelines is a necessary criterion for business implementation.

The issue of ethical AI is also a cultural issue. The nominal concept interpretation of ethics can be affected via different channels (individuum / organisation / society, religion, others). Establishing principles based on organisational values and integrated into the strategy (Code of Ethical AI) can promote the trustworthiness of AI systems and generate confidence between the parties involved. To achieve this, ethical AI must be organisationally interpreted and shaped according to experience, which will continuously improve via "self-governance" in response to social feedback and approach the set of harmonised ethical requirements. The author emphasises the development and application of ethical AI systems can be regarded as an ongoing iteration, which can create the theoretical principles of trustworthiness. Consequently, the author suggests organisations should reconsider ethical principles, *Figure 2* may serve as assistance.

Figure 2
Organisational interpretation of ethics - in author's view



Source: Own design

ANNEXES

Annex 1
Research questions and answers - summary of interviews

#	Q1	Q2	Q3	Q4
1.	top-down	positive, supportive attitude	group of modelling data and questions of data collection; no discrimination, encryption of customer data	project participants' involvement is high, supplier provides knowledge (workshop)
2.	top-down	varied, depends on AI skills	moral decisions of self-driving car, legal issues of data collection	nil, due to lack of time; training is too general
3.	top-down	positive	ethical issues in infancy; operator's responsibility	no training; system documentation is handed over when system is delivered
4.	top-down	mostly positive, depends on AI skills	ethical consideration did not arise during planning	there is
5.	bottom-up	positive	ethical consideration did not arise during planning	nil
6.	supplier→ top down, bottom up	support in words, refusal in acts	rather limited appearance: discriminative decision-making, inclusion	nil due to cost control
7.	top-down	positive, but limited openness (protection of data assets)	rather limited appearance: discriminative decision-making	competence centre being built at organisational level
8.	bottom-up	no or limited support	ethical consideration did not arise during planning	nil
9.	bottom-up	openness at beginning (diminished as costs were estimated)	ethical consideration did not arise during planning	nil
10.	top-down	positive supportive approach	visual display, security considerations	supplier provides
11.	supplier → top-down	cannot refuse because of executive pressure (can be made up for)	ethical consideration did not arise during planning	workshop held and documentation handed over when system is delivered
12.	top-down	positive, but parent company may be restrictive	in banking sector: security issues, data management	architects launched lectures for management
13.	top-down	positive supportive approach	ethical consideration did not arise during planning	regular corporate training sessions supported by IT, talks, brainstorming

Source: Own design

Annex 2
Regulations, guidelines used and relevant sections

Regulation	Title / chapter / article	Requirement	Summary content
Regulation (EU) 2019/1020 on Market Surveillance and Compliance of Products*	Title 1, Article 3, Section 19 (General provisions)	product presenting a risk	means a product having the potential to affect adversely health and safety of persons in general, health and safety in the workplace, protection of consumers, the environment, public security and other public interests protected by the applicable Union harmonisation legislation, to a degree which goes beyond that considered reasonable and acceptable in relation to its intended purpose or under the normal or reasonably foreseeable conditions of use of the product concerned
COM (2021) 206 final	Title I, Article 3, Section 1 (definitions)	concept of artificial intelligence	...software that is developed with [specific] techniques and approaches [listed in Annex 1] and can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with
	Title III, Chapter 2, Article 13 Annex III (high-risk AI systems)	Amendments to Annex III criteria to be applied to assess wdamage caused	(a.) The intended purpose of the AI system, (b) the extent to which an AI system has been used or is likely to be used, (c) the extent to which the use of an AI system has already caused harm [...] or adverse impact [...] as demonstrated by reports or documented allegations, (d) the potential extent of such harm or such adverse impact (group of persons affected); (e) the extent to which potentially harmed or adversely impacted persons are dependent on the outcome produced with an AI system, (g) the extent to which the outcome produced with an AI system is easily reversible; (h) the extent to which existing Union legislation provides for [...] effective measures to prevent or substantially minimise those risks
	Title III, Chapter 2, 8-15 Article	Requirements for high-risk AI systems	(8.) Compliance with the requirements: (9) A risk management system shall be established, implemented, documented and maintained; (10) Data and data governance; (11) Technical documentation; (12) Record-keeping; (13) Transparency and provision of information to users; (14) Human oversight; (15) Accuracy, robustness and cybersecurity
	Title VIII, Chapter III, Article 65 (1-9)	Procedure for dealing with AI systems presenting a risk at national level	Where the market surveillance authority of a Member State finds that an AI system does not comply with the requirements and obligations laid down in this Regulation, it shall without delay require the relevant operator to take all appropriate corrective actions to bring the AI system into compliance or to withdraw the AI system from the market
Ethical guidelines on trustworthy AI	Chapter II.	1-7.	Requirements of a trustworthy AI system: (1.) human agency and human oversight; (2.) technical robustness and safety; (3.) privacy and data governance; (4.) transparency; (5.) diversity, non-discrimination and fairness (6.) environmental and social well-being; (7.) accountability.
ALTAI	Full document	Full document	Elements of the assessment list: (1) Human Agency and Oversight (2) Technical Robustness and Safety (3) Privacy and Data Governance (4) Transparency (5) Diversity, Non-discrimination and Fairness (6) Environmental and Social well-being (7) Accountability

*Note: *on Market Surveillance and Compliance of Products*, and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011*
Source: Own design

Annex 3
Summary Table of the philosophy and ethics of philosophers referred to in the gap analysis

Philosopher	Philosophy	Ethics
	Greco-Roman philosophy Pre-Socrates	
Parmenides (515 BC – 470 BC)	monism, to be or not to be concept: separation of dimensions of being (permanent, eternal truth), or non-being (changing, transient); through intellectual (former) or sensory (latter) perception	ethics standards are parts of the eternal and permanent dimension, but they are expressed in the dimension of non-being, in the changing and transient world one can see
Socrates (469 BC – 399 BC)	central components are knowledge and moral, experience (argumentation and dialogue are important); intellectual thinking is a kind of absolute knowledge/truth about the world to be achieved by recognising human limitations (ignorance)	his ethics is built on knowledge and virtues; they can be improved by learning; intellect and related ethical behaviour based on argumentation and thinking
Xenophon (434 BC – 355 BC)	Intellectual thinking based on reasonable argumentation, logic and deduction is of key importance for correct decision-making	ethical behaviour is based on correct decisions by intellectual thinking and Socratic virtues
Platon (427 BC – BC)	dualistic concept (body and soul separated, philosophy of mind), two-world theory (world of existence is permanent and can only be accessed by the intellect)	ethical behaviour is based on: knowledge, justice and virtues learning helps correct decision-making (it must serve the social good)
Aristoteles (384 BC – 322 BC)	dualistic theory, philosophy of mind, experimental learning (perception); passive/active intellect; logical thinking and contemplation to understand importance of Socratic virtues	intelligent man is capable to understand truth, to use reason and act according to ethical values (social usefulness)
	Middle Age philosophy	
Saint Augustine (354 to 430)	contrary to bases in ancient Greece (intellect, logic, knowledge), faith is the bridge between the sensual world and the world of reason; patristic philosophy: understanding and applying divine truth (scriptures)	is rooted in Christian ideals and ethical conduct (approaching human happiness and God); free will and related responsibility
St Thomas Aquinas (1225–1274)	dominant scholastic philosophy in late Middle Ages: harmony of natural and religious truths, following argumentation and rationality - Aristotelian foundations; intellect to serve understanding of divine truth	virtues linked to faith in God; connection of free will and responsibility; morality is based on respect of human nature and is created by intellect promoting social well-being

William Ockham (1287–1347)	scholastic philosophy, universal concepts are nominal, mere mental constructions that do not exist as objective reality; criticism and scepticism, principle of simplicity is important in experiential learning while assumptions must be minimised	concepts related to ethics standards do not exist in themselves, they can be interpreted subjectively linked to persons and events, ethical values are the result of social convention and do not exist by themselves
Péter Pázmány (1570-1637)	Aristotelian bases but elements of scholastic philosophy: analysis of harmony of intellect (understanding the world) and faith in God (deep understanding if life)	strive for a happy life based on ethical values (to be found in the divine and related order of man)
	New Age philosophy	
René Descartes (1596–1650)	rationalism; return to bases: analysis of complex thinking processes and the operation of the mind; use of scientific methods helps understand objective reality; there is also a non-material (spiritual) world through inner experience of cognition)	two elements of his ethics: freedom and self-determination (conscience-based action); moral duty structures life; output (cause and effect) is not always unambiguous; rational decision-making and accepting consequences
Gottfried Wilhelm Leibniz (1646–1716)	human intellect (sense) can help understand higher knowledge (God's intention) contemplates automation of knowledge; "monad" (a building block of universe, a closed system that reflects the order of the world) is central to his views	composite theory: division of universal good into three parts: metaphysical, moral and physical good based on monads: individual and social happiness are in harmony, cooperation for the bigger good
Francis Bacon (1561–1626)	establishment of scientific methods; science can be based on empirical (objective) observation	society is responsible for setting up morals (education and knowledge of nature)
Thomas Hobbes (1588–1679)	theory of social contract (partial limitation of freedom); the natural state leads to chaos because of human egotism, so it is necessary to protect community (order, safety)	morality and ethical values are the result of social conventions; observation of laws and norms reflect ethical behaviour
John Locke (1632–1704)	knowledge is rooted in experience based on data acquired by the senses (no a-priori knowledge); the state is responsible for protecting people's rights	tolerance (free exercise of own rights without violating others'), individual freedom and self-determination have emphasis
David Hume	intelligence develops from experience perceived by the senses (role of sceptic) to understand the laws of the world	morality rooted in emotions is subjective and relative; ethical standards are conventions for living in society
Jean-Jacques Rousseau (1712–1778)	theme of social contract: people are basically happy until social restraints, classes and hierarchies deprive them - return to a natural state	compassion empathy and altruism are important, social conventions strive to suppress them; respect for individual freedom

	Modern New Age philosophy	
Immanuel Kant (1724–1804)	critical philosophy and general (social) will (ethical standards and laws); autonomy (action of the individuum based on reason)	moral action means observing the laws defined by human reason
John Stuart Mill (1806–1873)	utilitarianism: protection of individual freedom (by laws and social norms), democracy - governance in people's interest	actions assessed by their consequence and results (maximise social usefulness)
	Contemporary philosophy	
Daniel Dennett (1942–	the mind, consciousness and free will should be studied via sciences; thoughts and experiences are the outcome of complex cognitive processes, which can be distorted with cognitive limitations and errors; independent decision-making ability (defined by algorithms)	ethics standards are the outcome of social agreements varying with cultures, they can change in time; ethical decision-making based on reason and free will influences the creation of social values
Martha Nussbaum (1947–	the rights and dignity of the individual must be respected and protected by society irrespective of the individual's status, actions or experiences	empathy and moral sensibility are indispensable elements of human well-being, which relate to the protection of human dignity
Nick Bostrom (1973–	analyses the impact of scientific-technological development: principle of anti-realism (the perceived world is but a distorted partial image of reality); principle of trans-humanism (enhancing human potential)	safety brakes and guarantees must be established and built into the process of AI development; it is of vital importance to avoid catastrophic consequences

REFERENCES

Anderson, M. – Anderson, S. L. (2011): *Machine Ethics*. Cambridge, MA: Cambridge University Press, https://doi.org/10.1017/CBO9780511978036.

Anderson, S. L. (1995): Being Morally Responsible for an Action Versus Acting Responsibly or Irresponsibly. *Journal of Philosophical Research,* 20, 453–62, https://doi.org/10.5840/jpr_1995_10.

Asimov, I. (1942). Runaround. *Astounding Science Fiction*, March 1942. Online: https://web.williams.edu/Mathematics/sjmiller/public_html/105Sp10/handouts/Runaround.html. Magyarul: Körbekörbe. In Asimov, I.: *Robottörténetek*, http://users.atw.hu/asimov/downloads/Encyclopedia%20Galactica/01.%20k%C3%B6tet%20-%20Encyclopedia%20Galactica/Isaac%20Asimov%20-%20Robott%C3%B6rt%C3%A9netek.pdf.

Asimov, I. (1950). I, Robot. *Gnome Press*. Online: http://ekladata.com/-Byix64G_NtE0xI4A6PA1--01Hc/Asimov-Isaac-I-Robot.pdf. Magyarul: Asimov, I. (2019): Én, a robot (fordította: Békési József, Vámosi Pál). Budapest: GABO.

Awad, E. – Dsouza, S. – Kim, R. –Schulz, J. – Henrich, J. – Shariff, A. – Bonnefon, J-F. – Rahwan, I. (2018): The Moral Machine Experiment. *Nature,* 563(7729), 59–64, https://doi.org/10.1038/s41586-018-0637-6 .

Beijing Academy of Artificial Intelligence (2019): Beijing AI Principles. Datenschutz und Datensicherheit 43, 656. Online: https://www.semanticscholar.org/paper/Beijing-AI-Principles/a703873b9c697c05b9146a5df790745b6f303857.

Cointe, N. – Bonnet, G. (2016): Ethical Judgment of Agents' Behaviors in Multi-Agent Systems. *AAMAS*, 1106–1114.

Dennett, D. – Fleig-Goldstein, B. –Friedman, D. (2019): Dennett Explained: An interview with Daniel Dennett. *ALIUS Bulletin,* 3, 11–25, https://doi.org/10.34700/7gkw-zh08.

Dennett, D. (1984): Cognitive Wheels: The Frame Problem of AI. *Minds, Machines, and Evolution*. Cambridge: Cambridge University Press, 129–152. Online: https://www.researchgate.net/publication/225070451_Cognitive_Wheels_The_Frame_Problem_of_AI.

Dennett, D. (2019): What can we do? In Brockman, J. (ed.) (2019): *Possible Minds: 25 Ways of Looking at AI* (Chapter 5). Online: https://ase.tufts.edu/cogstud/dennett/papers/What_Can_We_Do.pdf.

Denning, P. J. – Denning, D. E. (2020): Dilemmas of artificial intelligence. *Communications of the ACM*, 63(3), 22–24, https://doi.org/10.1145/3379920 .

Dreyfus, H. L. (1972): *What computers* still *can't do: A critique of artificial reason*. Cambridge, MA: MIT Press. Online: https://terrorgum.com/tfox/books/whatcomputersstillcantdo_acritiqueofartificialreason.pdf.

Dreyfus, H. L. (2007): Why Heideggerian AI Failed and how Fixing it would Require making it more Heideggerian. *Artificial Intelligence,* 171(18), 1137–1160, https://doi.org/10.1016/j.artint.2007.10.012.

Drucker, F. P. (2001): What is business ethics? *The Public Interest*, 35, 18–36. Online: https://edisciplinas.usp.br.

European Banking Authority (2021): EBA Discussion Paper on Machine Learning for IRB Models, https://www.eba.europa.eu/.

European Banking Federation (2019): EBF position paper on AI in the banking industry, https://www.ebf.eu.

European Commission (2018): COM(2018) 237 final. Online: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237&from=EN.

European Commission (2019): Ethics Guidelines for Trustworthy AI. Online: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

European Commission (2020a): The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment. Online: https://op.europa.eu/en/publication-detail/-/publication/73552fcd-f7c2-11ea-991b-01aa75ed71a1/language-en/format-PDF/source-286827461.

European Council (2020b): Special European Council meeting (1–2 October 2020) – Outcome, EUCO 13/20, 2020, 6, 4. Online: https://www.consilium.europa.eu/en/meetings/european-council/2020/10/01-02/.

European Commission (2021): Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts. COM(2021) 206 final. Online: https://eur-lex.europa.eu.

European Commission (2022): New approach to enable global leadership of EU standards promoting values and a resilient, green and digital Single Market. Online: https://ec.europa.eu/commission/presscorner/detail/en/ip_22_661.

European Parliament and the Council (2019): Regulation (EU) 2019/1020 of the European Parliament and of the Council of 20 June 2019 on market surveillance and compliance of products and amending Directive 2004/42/EC and Regulations (EC) No 765/2008 and (EU) No 305/2011 (Text with EEA relevance.) Online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELLAR:903d90ee-9712-11e9-9369-01aa75ed71a1.

Executive Office of the President National Science and Technology (2016): Preparing For The Future Of Artificial Intelligence. Online: https://obamawhitehouse.archives.gov/sites/default/files/whitehouse_files/microsites/ostp/NSTC/preparing_for_the_future_of_ai.pdf.

Floridi, L. – Chiriatti, M. (2020): GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds & Machines*, 30, 681–694, https://doi.org/10.1007/s11023-020-09548-1.

Floridi, L. – Cowls, J. (2019): A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review,* https://doi.org/10.1162/99608f92.8cd550d1 .

Friedman, B. – Nissenbaum, H. (1996): *Bias in Computer Systems*. In Weckert, J. (ed.) *Computer Ethics* (Chapter 20.). London: Routledge, https://doi.org/10.4324/9781315259697-23 .

Hagendorff, T. (2020): The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*, 30, 99–120, https://doi.org/10.1007/s11023-020-09517-8 .

Héder, Mihály (2020): *Mesterséges intelligencia – Filozófiai kérdések, gyakorlati válaszok* [Artificial intelligence – philosophical questions, practical answers]. Budapest: Gondolat Kiadó. ISBN: 9789635560509.

Institute of Electrical and Electronics Engineers (2016): The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. Online: https://standards.ieee.org/wp-content/uploads/import/documents/other/ec_about_us.pdf.

Institute of Electrical and Electronics Engineers (2019): Ethical Aspects of Autonomous and Intelligent Systems. Online: https://globalpolicy.ieee.org/wp-content/uploads/2019/06/IEEE19002.pdf.

Institute of Electrical and Electronics Engineers (2021): The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems Industry Connections Activity Initiation Document (IC-AID). Online: https://standards.ieee.org/wp-content/uploads/import/governance/iccom/IC16-002-Global_Initiative_for_Ethical_Considerations_in_the_Design_of_Autonomous_Systems.pdf.

Jobin, A. – Lenca, M. – Vayena, E. (2019): The Global Landscape of AI Ethics Guidelines. *Nature Machine Intelligence*, 1, 389–399, https://doi.org/10.1038/s42256-019-0088-2 .

Kirkpatrick, J. (2015): Drones and the Martial Virtue Courage. *Journal Of Military Ethics*, 14(3–4), 202–219, https://doi.org/10.1080/15027570.2015.1106744 .

McCarthy, J. – Minsky, M. L. – Rochester, N. – Shannon, C. E. (1955): A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. *AI Magazine*, 27(4), 12, https://doi.org/10.1609/aimag.v27i4.1904.

Von Neumann, J. (1958): *The Computer and the Brain*. Yale University Press, Inc. Online: https://complexityexplorer.s3.amazonaws.com/supplemental_materials/5.6+Artificial+Life/The+Computer+and+The+Brain_text.pdf.

Von Neumann, J. (1963): The General and Logical Theory of Au-tomata. Online: https://www.semanticscholar.org/paper/The-General-and-Logical-Theory-of-Au tomata-Neumann/e8538f11920fa6e56b3d34771bb330bd3e07281d.

Nilsson, N. J. (2010): *The Quest For Artificial Intelligence. A History of Ideas and Achievements*. Cambridge University Press. https://doi.org/10.1017/cbo9780511819346.

OECD (2019): Recommendation of the Council on Artificial Intelligence. Online: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

Prisznyák, Alexandra (2023a): Mesterséges intelligencia a bankszektorban [Artificial intelligence in the banking sector]. Book review. *Economy and Finance*, 9(3), 333–339, https://bankszovetseg.hu/Public/gep/333-340%20E%20Kiss_et_al_konyv.pdf.

Prisznyák, Alexandra (2023b): Philosophical Questions of the Manifestation of Natural Intelligence (book review). *Financial and Economic Review*, 23(1), 164–168. Online: https://enhitelintezetiszemle.mnb.hu/letoltes/fer-22-1-br2-prisznyak.pdf.

Reiter, E. – Dale, R. (2021): Building Applied Natural Language Generation Systems. *Natural Language Engineering* , 27(1) , 113–118, https://doi.org/10.1017/s1351324997001502 .

Searle, J. R. (1980): Minds, Brains and Programs. *Behavioral and Brain Science*, 3(3), 417–424, https://doi.org/10.1016/b978-1-4832-1446-7.50007-8 .

Sejnowski, T. J. (2023): Large Language Models and the Reverse Turing Test. *Neural Computation*, 35, 309–342, https://doi.org/10.1162/neco_a_01563.

The Guardian (2020): A robot wrote this entire article. Are you scared yet, human? Online: https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3.

Török, Bernát – Ződi, Zsolt (2021): *A mesterséges intelligencia szabályozási kihívásai – Tanulmányok a mesterséges intelligencia és a jog határterületeiről* [Regulatory challenges of artificial intelligence – Studies in the border area of artificial intelligence and law]. Budapest: Ludovika Egyetemi Kiadó. ISBN: 9789635314836.

Turing, A. M. (1950): Computing Machinery and Intelligence. *Mind*, 59(236), 433–460, https://doi.org/10.1093/mind/lix.236.433.

UNESCO (2020): Recommendation on the Ethics of Artificial Intelligence. Online: https://unesdoc.unesco.org/ark:/48223/pf0000380455.

Veruggio, G. (2007): *The EURON roboethics roadmap*. 2006 6th IEEE-RAS International Conference on Humanoid Robots (Red Hook, NY; Genoa: IEEE), 612–617, https://doi.org/10.1109/ichr.2006.321337.

Wang, P. (2019): On Defining Artificial Intelligence. *Journal of Artificial General Intelligence*, 10(2), 1–37, https://doi.org/10.2478/jagi-2019-0002 .

Weizenbaum, J. (1976): *Computer power and Human Reason: From Judgment to Calculation*. New York: W. H. Freeman and Company. Online: http://blogs.evergreen.edu/cpat/files/2013/05/Computer-Power-and-Human-Reason.pdf.

Wiener, N. (1948): *Cybernetics or Control and Communication in the Animal and the Machine*. Cambridge, MA: The MIT Press. https://doi.org/10.7551/mitpress/11810.001.0001.

Yu, H. – Liu, Z.; Wang, Y. – Jiang, X. (2018): Building Ethics into Artificial Intelligence. *IEEE Intelligent Systems*, 33(4), 77–83. https://doi.org/10.24963/ijcai.2018/779 .

Sources used for part on philosophy:

https://plato.stanford.edu/index.html