SEKYRA FOUNDATION
Ethics & Bioethics (in Central Europe), 2026, 16 (3–4), 146–164 DOI:10.2478/ebce-2026-0014

The ethics of regulation: Social contract insights on the 2024 European Union Artificial Intelligence Act

Leandro Loriga[1]

Abstract
The paper provides a critical analysis of the EU AI Act (Regulation 2024/1689) within the broader context of contemporary AI developments. Starting from an historical overview on the development of advanced AI systems, it moves the focus onto the intrinsic meaning of Artificial Intelligence to highlight how, despite such fascinating wording, there cannot be a shift of responsibility onto the systems themselves—as was proposed, for example, by the European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (EUR-Lex - 52017IP0051); not until, at least, singularity is achieved. Moral and legal responsibility, therefore, lies with planners, developers, implementers, and all other stakeholders who have an interest in AI systems. Through the lens of social contract theory, drawing on Hobbes' *Leviathan* and recent AI ethics scholarship, the paper identifies several weaknesses within the EU AI Act itself, including the rigidity of its risk-based model, its focus on technical compliance rather than societal trust, and gaps in governance and accountability. Building on these findings, it proposes potential solutions, including dynamic risk assessment, stronger participatory mechanisms, and clearer enforcement structures. The need thus emerges to recognize and acknowledge a collective type of responsibility arising from a social contract among all parties involved. The ideal scope is to scale back individual interests in favor of the collective good. While the Act marks the first attempt to create a binding moral and legal framework for trustworthy AI among European Union Member States, its ultimate success will depend on its flexibility and on cultivating a shared sense of responsibility and partnership among all stakeholders.

Keywords: Artificial Intelligence (AI); AI Ethics; Social Contract Theory; EU AI Act 2024; AI Regulation; Ethical Governance; AI Accountability; AI Risk Assessment; Trustworthy AI; AI and Society; High-Risk AI Systems; AI Transparency; AI Governance Framework

Introduction

Artificial intelligence (AI) is now more than ever under the spotlight of academic, corporate, and lay audiences. What once was considered a far-fetched idea, that is, the creation of a non-biological type of life with characteristics resembling or surpassing those of humans, is now getting closer by the day. The idea of human-made artificial entities and their possible implications has captivated a wide variety of audiences for a long time.

Samuel Butler, for example, in his fictional work *Erewhon* (Butler, 1987), develops an allegorical reflection on his contemporary Victorian England through the analysis of an imaginary society, which he uses to echo topics such as religion, justice, and the era's general attitudes. Drenched in the scientific innovations of his time, he critically touched technological advancements (Butler, 1987, pp. 224–260) by intertwining new evolutionary ideas—such as Darwin's *On the Origin of Species*, of which, however, he later became a mild critic (Pauly, 1982)—with the machines of the industrial revolution. He used this to highlight concerns about dependence on technology, blind faith in scientific progress, and the fear of losing control over artificial creations. On a more explicit note, Thea von Harbou (2015) presents a dystopian future in Metropolis, in which technological innovation is ruled by a purely capitalist logic and used to oppress and divide people. It is, however, perhaps with the work of Isaac Asimov that AI fully caught the public's attention. His influential *Three Laws of Robotics* (Asimov, 2014, p. 1) appear as a well-established artifact in popular culture. Its influence in discussions of the ethics and morality of AI systems seems to overshadow more specialized approaches to decision-making, especially in sensitive areas such as healthcare, as might the four bioethics principles

[1] Masaryk University (Czech Republic); email: leandro.loriga@med.muni.cz; ORCID: 0000-0002-8899-6847

of Beauchamp and Childress (Beauchamp & Childress, 2013) for moral and ethical decision-making. It is also true that none of Beauchamp and Childress' work was ever put on the big screen in a way like Alex Proyas and Will Smith did with *I, Robot*.

These are just a few among many examples that foreshadow today's concerns. Such issues include the decline in human cognitive abilities from excessive screen exposure and technology reliance (see, for example, *digital dementia* Sandu & Nistor, 2021; Spitzer, 2012), use of autonomous AI weapons without proper safeguards (Christie, Ertan, Adomaitis, et al., 2024), algorithm bias in AI analysis (Leavy, O'Sullivan & Siapera, 2020; Walker, Dillard-Wright & Iradukunda, 2023) that could endanger human rights, and the idea of AI Singularity (Jalšenjak, 2020).

The beginning of the history of AI in the western world can symbolically be positioned with the Dartmouth Summer Research Project of 1956 (Cordeschi, 2010, pp. 155–174; Kline, 2011; Moor, 2006) where a pool of experts aimed to pin-point among others, the characteristics of human reasoning in a precise enough fashion to be simulated artificially with the aid of machines. Around a decade later, Joseph Weizenbaum developed the ELIZA Chatbot (Berry, 2023; Sharma, Goyal & Malik, 2017; Shrager, 2024; Weizenbaum, 1966), a program that laid the foundation of Natural Language Processing (NLP) and opened the gate to machine-human interactions, highlighting the dangers of emotional attachment. This ELIZA effect emerged through the DOCTOR experiment, in which ELIZA emulated a Rogerian psychotherapist by reflecting users' statements back to them (for a detailed overview, see Dillon, 2020; Natale, 2021).

It is undeniable that AI research and application have been slowly but steadily expanding across public and private domains, including healthcare. Early rule-based expert systems (*if-then rule*s), such as the 1970s MYCIN project for diagnosing bacterial infections (see Buchanan & Shortliffe, 1985), paved the way for AI-powered CAD systems (Hassan, Hamad & Mahar, 2022). New neural network approaches, such as Deep Learning, can now analyze raw data without human aid (Malliori & Pallikarakis, 2022; Wang, 2024), and predictive tools such as IBM Watson, DeepMind, and AlphaFold solutions (De Fauw, Ledsam, Romera-Paredes et al., 2018; Powles & Hodson, 2017; Tupasela & Di Nucci, 2020) are drastically changing the way healthcare is understood and implemented today, while simultaneously shifting the divide between therapy and enhancement. Such advances in healthcare diagnosis, prognosis, and prediction are being combined with widely available AI tools that are becoming a regular part of daily life. The launch of Apple's Siri, Amazon's Alexa, and Google's Assistant in the early 2010s popularised basic AI models for everyday use. It paved the way for the widespread impact of significantly more advanced models, including OpenAI's ChatGPT, Anthropic's Claude, Google's Gemini, and many others. AI now seems like a daily tool that not only enhances task efficiency but can also generate outputs (consider, for example, the popularity of OpenAI's DALL·E Series or MidJourney). The widespread influence of AI is undeniable, and public participation is contributing to this new AI revolution (Marmolejo-Ramos, Workman, Walker et al., 2022; Seth, 2024; Yigitcanlar, Degirmenci & Inkinen, 2024). Any fast-paced change, however, necessarily raises concerns, and with it the need to step back and reflect critically on current events to avoid being swept away by the waves of novelty, optimism, and enthusiasm.

This paper begins by offering an overview of key issues surrounding AI today, starting with the intrinsic challenges of categorizing artificial systems as intelligent. It then shifts the focus to the proliferation of ethical guidelines for AI systems and models, leading to a patchwork of ad hoc solutions. This fragmented approach shows the pressing need to recognize the participatory nature of morality and to establish a clear, enforceable social contract to guide AI development and ensure accountability for its future trajectory. Such a direction is not only necessary but also promising, given the recent adoption of the EU AI Act, the first legislation

of its kind to have binding legal authority for European Union Member States. The paper concludes by examining the Act in detail, identifying three potentially contentious areas—risk assessment, trustworthiness, and governance—and suggesting practical steps for its effective implementation.

Intelligence in the artificial

As time changes, so does collective sensibility on certain matters. For some, we are currently living through the fourth industrial revolution (see, for example, Ross & Maynard, 2021; Schwab, 2016), characterized by an exponential, overreaching scope of technology in daily lives. AI is leading to such a change. "AI is the field devoted to building artifacts capable of displaying, in controlled, well-understood environments, and over sustained periods of time, behaviours that we consider to be intelligent, or more generally, behaviours that we take to be at the heart of what it is to have a mind" (Arkoudas & Bringsjord, 2014, p. 34)

The idea *that AI operates only in well-defined environment*s is becoming obsolete, since today AI models are deeply intertwined with complex sociocultural contexts, as in the case of autonomous vehicles (Geisslinger, Poszler, Betz et al., 2021, 2022; Keeling, 2020; Paulo, 2023; Zhan & Wan, 2024). Similarly, the mention of *intelligence* hints at a type of behavior one might recognize as intelligent, even though the entity exhibiting it (whether human or otherwise) might not be intelligent at all. Of course, a starting position would be to define what it means to be intelligent, especially within the AI debate (see, for example, Gignac & Szodorai, 2024; Wang, 2019). Take, for example, a GPS system that can drive from point A to point B and can similarly adjust the itinerary along the way based on newly processed inputs, such as unexpected roadblocks, predicted traffic jams at specific intersections, and similar. Could such software be considered intelligent?

> Intelligence is a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience. It is not merely book learning, a narrow academic skill, or test-taking smarts. Rather, it reflects a broader and deeper capability for comprehending our surroundings "catching on," "making sense" of things, or "figuring out" what to do (Gottfredson, 1997, p. 13).

This raises the question of whether machine and human processing are comparable—not just in power, but also in quality and interpretation—or whether imitation is truly the same as genuine behavior. One answer to such a putative question is a straightforward no; the fact that someone or something can mimic something else does not mean this someone or something understands what is being mimicked beyond achieving a specific task, fulfilling a specific requirement, or obtaining a specific response. In other words, to recognize something as expressing intelligent behavior, that something needs to ponder on the intrinsic meaning of such a behavior, rather than on the extrinsic one. This is to say that meaning, perhaps, lies in the syntax of things rather than in the end result.

This and similar accusations have been moved toward the machine world (see, for example, Floridi et al., 2009) despite the fact that recent AI technologies might very well have a high success rate in fooling humans by pretending to be human themselves (Jones & Bergen, 2023, 2024; Rathi, Taylor, Bergen et al., 2024). For some, therefore, AI, rather than having intelligence, provides intelligent-like responses without being intelligent itself; "a reservoir of smart agency on tap" (Floridi, 2019, p. 3). For this and similar positions to be considered valid, however, one would need to first identify the putative difference between the inputs humans receive during their lives and those advanced AI models receive during training. Similarly, one would need to identify differences in how such inputs are processed and interpreted in a fashion accurate enough to state that no commonalities exist between humans and machines. Such

putative differentiation might be challenged by leveraging topics such as Black Box outputs for opaque decision-making (for example, Von Eschenbach, 2021), namely, the apparent inexplicability of certain conclusions that are not easily retractable, as in simpler if-then rule models. Similarly, unsupervised Machine Learning Algorithm (for a review see Alloghani et al., 2020) could be used to argue that the information pool used by AI models for learning, and the way such a pool is exploited, might mimic the information pool and approach used by humans for learning (imagine a very small child let free to explore the world without strict guidelines and reinforcement methods). A more difficult account to oppose would be one that relies on elements commonly included in definitions of intelligence that go beyond logical capabilities, such as moral compass and emotional intelligence (Greene, Nystrom, Engell et al., 2004; Helion & Pizarro, 2015; Mestvirishvili, Mestvirishvili, Kvitsiani et al., 2020). Even if this is true for now, the divide may eventually blur, particularly in light of speculation about AI's potential to reach consciousness (see, for example, Butlin et al., 2023; Juliani et al., 2022; Lopes, 2023).

Although this and similar topics require careful consideration, they fall outside the scope of the current work and will not be dwelt on here. Nonetheless, it was necessary to highlight some of the challenges inherent to the concept of intelligence to clear the ground for possible disputes. With these challenges addressed—and given that machines are not yet intelligent beings in the human sense—it is now possible to turn to some of the most pressing issues with today's AI guidelines. The following section will do so and will serve as a runway to introduce the EU AI Act.

From past to present: Ethical challenges of AI guidelines

Over the years, many guidelines for AI behavior have emerged to set boundaries on what is morally permissible. For example, Floridi and Cowls (2019) review six major frameworks developed by reputable sources. Yet one risk of having so many contributors is the uncontrolled spread of *ad hoc* AI ethics guidelines, created by independent stakeholders primarily to signal moral responsibility and present their service offering as ethically sound. Such a situation makes the very purpose of having ethical guidelines for the development and use of AI void. This is to say that such a variety of ethical guidelines will turn—and is already turning—into a shopping spree for the best ethical-principle-toolset fit for the purpose:

> [T]he malpractice of choosing, adapting, or revising (mixing and matching) ethical principles, guidelines, codes, frameworks, or other similar standards (especially but not only in the ethics of AI), from a variety of available offers, in order to retrofit some pre-existing behaviours (choices, processes, strategies, etc.), and hence justify them a posteriori, instead of implementing or improving new behaviours by benchmarking them against public, ethical standards (Floridi, 2019, p. 186).

Such a situation will necessarily lead to a loss of significance due to the proliferation of meanings and perspectives; if anyone has the right to say anything about everything in an uncontrolled, qualitatively equal manner, then meaning is lost (see, for example, Cohen, 1993). In such a situation, the idea of *explicability*, for example (Floridi & Cowls, 2019), becomes a necessity. "The addition of the principle of 'explicability,' incorporating both the epistemological sense of 'intelligibility' (as an answer to the question 'how does it work?') and in the ethical sense of 'accountability' (as an answer to the question 'who is responsible for the way it works?'), is the crucial missing piece of the AI ethics jigsaw" (Floridi & Cowls, 2019, pp. 8–9).

Herrera and Calderón (Herrera & Calderón, 2025) expand on this debate by introducing the LoBOX governance ethic, a framework that treats opacity as a governable feature through role-sensitive explanations and institutional trust, rather than as a flaw to be eliminated. Similarly,

Hagendorff (2020) highlights the issue of *pro forma* rather than enforceable guidelines. By referencing the ethical decision-making in software development (McNamara, Smith & Murphy-Hill, 2018), the author states that "the effectiveness of guidelines or ethical codes is almost zero and that they do not change the behavior of professionals from the tech community" (Hagendorff, 2020, p. 108) and that

> AI ethics is failing in many cases. Ethics lacks a reinforcement mechanism. Deviations from the various codes of ethics have no consequences. Moreover, when ethics is integrated into institutions, it primarily serves as a marketing strategy. Furthermore, empirical experiments show that reading ethics guidelines has no significant influence on the decision-making of software developers (Hagendorff, 2020, p. 113).

What is needed, one might argue, is a shift toward a virtuous type of morality for conscious development of AI systems (Hagendorff, 2020, pp. 113–114). It would be counterintuitive to argue for anything other than enhanced moral awareness for all stakeholders involved in AI, including software developers, programmers, data scientists, machine learning engineers, marketing teams, executives, business leaders, and investors. It is also true, however, that such a statement can easily hold its value for any other type of human endeavor, which challenges the intrinsic value of such propositions.

The issue today, caused by the explosion of AI systems, is more pragmatic: AI models are now fully integrated into people's daily lives, meaning they are already here and reshaping how people live and perceive the world. Losing oneself in philosophical speculation on the goodness of certain elements and the badness of others is beside the point and obsolete. One might very well defend any approach to ethics and morality, be it a deontological one, a consequentialist one, or a virtue-based one, through semantic finesse and logical reasoning. This does not mean, however, that elaboration will go beyond the written text and the published paper toward practical application, especially when such dynamics are governed by capitalist logic in the corporate environment. What is needed is a practical agreement that serves the common good shared by the people, an enforcer to maintain it, and a well-established accountability mechanism.

A social contract for AI systems

The idea of framing AI within social contract theory is not new. Rahwan, for example, proposes the idea of expanding the various Human-in-the-Loop (HITL) approaches—that is, keeping the Human element as a supervising strategy to AI decisions—to a Society-in-the-Loop (SITL) (Rahwan, 2018) where the focus is moved from the individual to the interaction among the various stakeholders for the greater societal good.

> While HITL AI focuses on embedding the judgment of individual humans or groups into the optimization of AI systems with narrow impact, SITL focuses on embedding the values of society as a whole into the algorithmic governance of societal outcomes with broad implications. In other words, SITL becomes relevant when the scope of both the inputs and outputs of AI systems is very broad (Rahwan, 2018, p. 7).

While HITL is neither a new nor alien concept both for the public (see for example Bisen, 2022) nor for academic discourse (Amershi, Cakmak, Knox et al., 2014; Mosqueira-Rey, Hernández-Pereira, Alonso-Ríos et al., 2023) especially when it comes to medical diagnostic tools (see Chandler et al., 2022) due to the need to keep a level of human control to system which are far from being perfect while at the same time trying to demystify the inexplicability of certain algorithmic systems (Pasquale, 2015), SITL is somewhat new and seems to fall nicely within contemporary events.

Thomas Hobbes' *Leviathan lays* the basis for the social contract by arguing that without a common power, people exist in a perpetual state of war. "Hereby it is manifest that during the time men live without a common power to keep them all in awe, they are in that condition which is called war; and such a war as is of every man against every man" (Hobbes, 1997, p. 77). It is from such a state that the social contract emerges, born from the people's need to seek peace and collaboration for what Hobbes calls a *commodious* living. "The passions that incline men to peace are: fear of death; desire of such things as are necessary to commodious living; and a hope by their industry to obtain them" (Hobbes, 1997, p. 79). It is not the scope of the paper to dwell in details on social contract nor to its subsequent evolutions (Gauthier, 2006; Locke, 2010; Rawls, 2005; Rousseau, 2012; Skyrms, 2014), but rather to highlight how certain characteristics, through a lifting and shifting exercise, can benefit as well as help interpreting contemporary ethical AI guidelines models.

Before doing that, it is necessary to comment on the relationship between law and morality. The starting postulate is that we live in a society in which our livelihood depends on our neighbors. That is, we live in a mutually supportive system where, in one way or another, whether we like it or not, we look after each other, even if not explicitly and to the full extent of our possibilities, at least implicitly and to some degree. Such a mutually supportive system, which is today enforced by law, presupposes that what is good is caring for our fellow companions, be these direct neighbors or the faraway ones. What is bad—and depending on the situation, illegal—on the other hand, is to ignore other people's suffering and struggles, to ignore the wounded on the street after a car crash, or to omit aid in any reasonable enough type of situation that would not endanger our role as an accidental passerby and witness of other people's misfortune. If this is true (and it clearly is today), then we must necessarily see morality as an intrinsic, mutually supportive social human behavior. Consequently, safeguards are needed to maintain such a socially moral and righteous situation. The best way to do that is through enforcement methods, such as the legal system, with a notable mention also to Cesare Beccaria, Italian Enlightenment philosopher and jurist, and his *Dei delitti e delle pene* (Beccaria, 2018), in which he advocates the necessity of a solid legal system and the certainty of punishment for committed crimes.

A good starting point is to note that, although much of the world follows a democratic model, some aspects remain comparable to Hobbes' State of Nature—especially under the free-market capitalist model dominating the West. This shows why enforceable laws are necessary in AI ethics:

> It could be argued that the unrestrained appetite for power, freedom, and destruction displayed by modern corporations is reminiscent of the Hobbesian State of Nature … In the State of Nature, primitive man would create weapons and specific tools so that he could kill and destroy anybody that interfered with the advancement of his interests. Similarly, modern corporations also possess the will, incentive, and power, as well as a variety of tools and instruments, to overcome any obstacles to the achievement of their self-interests, including nature, competing firms, popular opposition, ideological antagonists, and government rules and regulations (Filip, 2020, pp. 324–325).

The idea of ruthless competition for resources and control is as present today as Hobbes speculated. The difference, however, is that this competition has been transposed to a different level of abstraction, where the main tools for control are no longer weapons, threats of violence, or brute force, but rather financial flows (Ho, 2005). This is clear enough. The next step is voluntary agreement; men must be willing to give up certain natural rights (*jus naturale*), including complete freedom, for peace and security. "That a man be willing, when others are so too, as far-forth as for peace and defence of himself he shall think it necessary, to lay down this right to all things; and be contented with so much liberty against other men, as he would

allow other men against himself" (Hobbes, 1997, p. 80). Such a position would translate for AI stakeholders into recognizing the value of adhering to specific ethical standards that prioritize social well-being over market edge and profitability. Caron and Gupta, for example, propose a "socially accepted purpose, a safe and responsible method, a socially aware level of risk involved, and a socially beneficial outcome" (Caron & Gupta, 2020). A practical example of such collective effort toward a mutual agreement is the Montréal Declaration on Responsible AI, which comprises 10 principles that emerged from collective discourse on AI use (2018).

> On November 3, 2017, the Université de Montréal launched the co-construction process for the Montréal Declaration for a Responsible Development of Artificial Intelligence (Montréal Declaration). A year later, the results of these citizen deliberations are public. Dozens of events were organised to stimulate discussion on social issues that arise with artificial intelligence (AI), and 15 deliberation workshops were held over three months, involving over 500 citizens, experts and stakeholders from all backgrounds (2018, loc. Context).

Such a voluntary agreement needs to be coordinated by a common power, that is, a central authority capable of enforcing it. "Covenants, without the sword, are but words and of no strength to secure a man at all" (Hobbes, 1997, p. 103).

Many critiques have been made of a self-regulated system governed by corporations. For example, Francés-Gómez (2023), beyond criticizing the clear lack of regulatory power emanating from AI developers themselves, who are, in any case, ruled by a capitalist market where profit is all that matters, calls for a binding legal framework, accountability, and an enforcement system at the global level. This means establishing and maintaining a cross-border accountability mechanism. "For the laws of nature, as justice, equity, modesty, mercy, and, in sum, doing to others as we would be done to, of themselves, without the terror of some power to cause them to be observed, are contrary to our natural passions, that carry us to partiality, pride, revenge, and the like" (Hobbes, 1997, p. 103).

For the present purpose, this would mean external audits conducted by independent review boards and similar entities and enforced by statutory powers such as those of the government and states. "Regulatory bodies should develop and implement clear policies regarding AI misuse. This could include periodic audits of AI usage, transparent reporting requirements for AI-assisted work, and the establishment of an ethics review board specialising in AI applications" (Lin, 2025, p. 2728).

A possible issue with the idea of a voluntary agreement, as presented by Caron and Gupta, is that it is presented as a *high-level, flexible framework* (Caron & Gupta, 2020, p. 4). Such a bird's-eye-view type of abstraction needs to be boiled down to practical, easily implementable guidelines, and the flexible framework needs to become more rigid. In other words, the principles that guide AI stakeholder behavior need to become enforceable rules.

Toward a common framework for AI governance: The EU - 2024/1689 - EN - EUR-LEX as the first social contract

The development of a specific moral framework for AI is a commendable goal but determining how to achieve it is the challenge. In the previous section, it was proposed that a possible way forward from this impasse is to focus on a social contract comprising at least three core elements: a social agreement, an enforcer, and an accountability mechanism. Mentxaka and colleagues (Mentxaka, Díaz-Rodríguez, Coeckelbergh et al., 2025) expand on this by stressing democracy as an essential dimension of trustworthy AI governance, presenting a dual taxonomy that identifies both the risks AI poses to democratic institutions and its potential contributions to transparency, participation, and fairness. This seems particularly fitting because, although we currently do not live in a State of Nature similar to the one speculated by Hobbes, there are

nonetheless certain areas that are more prone to resemble such a state (see, for example, the analogy between war and corporate environment, MacFarlane, 1999; Oneţ & Ciocoi-Pop, 2022). An important remark now must be made; all the above mentions about social contract and AI and the need for enforceable and strict guidelines precede one of the most important developments concerning AI in recent times; REGULATION - EU - 2024/1689 - EN - EUR-LEX (from now on EU AI Act), adopted on June 13, 2024 by the European Parliament, and subsequently published in the Official Journal of the European Union on July 12, 2024. The importance of such an Act is, above all, the establishment of a clear responsibility framework which cleanses past misconceptions. One such misconception is the attempt to establish Electronic Personhood for Liability and Responsibility purposes for AI models through the 2017 EUR-Lex - 52017IP0051

> creating a specific legal status for robots in the long run, so that at least the most sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties independently (European Union, 2017, para. 59F).

Even though such a proposition was clearly opposed and never went through (*Robotics Open Letter*, 2018), the EUR-Lex - 52017IP0051 laid the foundation for legally binding guidelines and a practical governance method of AI within the European Union, which are now finally being implemented through the EU AI Act (European Union, 2024).

> The purpose of this regulation is to improve the functioning of the internal market by laying down a uniform legal framework in particular for the development, the placing on the market, the putting into service and the use of artificial intelligence systems (AI systems) in the Union, in accordance with Union values, to promote the uptake of human centric and trustworthy artificial intelligence (AI) while ensuring a high level of protection of health, safety, fundamental rights as enshrined in the Charter of Fundamental Rights of the European Union (the 'Charter'), including democracy, the rule of law and environmental protection, to protect against the harmful effects of AI systems in the Union, and to support innovation... (European Union, 2024, recital. 1).

The document comprises 180 recitals that provide the foundation for the Act. Of particular interest are recital 1—which laid the foundation of the Act itself and highlights, right from the start, ideas such as *human-centric and trustworthy AI* (reiterated in Chapter 1 – General Provisions)—and recital 31, which introduces the topic of social scoring (further developed in Chapter 2, Article 5, para. 1c). "AI systems providing social scoring of natural persons by public or private actors may lead to discriminatory outcomes and the exclusion of certain groups. They may violate the right to dignity and non-discrimination and the values of equality and justice" (European Union, 2024, recital. 31). Particularly important for the present purpose is to consider the chapters that adopt a more morally charged tone, namely *Chapter 2: Prohibited AI Practices*, *Chapter 3: High-Risk AI Systems*, *Chapter 4: Transparency*, and *Chapter 7: Governance*.

Chapter 2, comprising Article 5 on *Prohibited AI Practices*, boils down core elements of unethical AI application, such as subliminal techniques and the exploitation of vulnerabilities to manipulate behavior, social scoring and predictive risk assessment, which might lead to bias and market edge, facial recognition databases, biometric categorization and identification, which instead hints at large-scale biometric data collection, storage, and processing of sensitive data. Chapter 3 comprises five sections of 43 articles and tackles high-risk AI systems. Among them are systems aimed at biometric identification and infrastructure management, such as

certain aspects of the upstream and downstream segments of the oil and gas industry, including automated drilling control systems and refinery process optimization, to mention a few. Similarly, the healthcare setting is also classified as a high-risk AI system (listed in Annex 3, para. 5). The chapter highlights those instances in which an AI system could endanger fundamental human rights. Interestingly, Article 9 mentions

> The risk management system shall be understood as a continuous iterative process planned and run throughout the entire lifecycle of a high-risk AI system, requiring regular systematic review and updating. It shall comprise the following steps… the estimation and evaluation of the risks that may emerge when the high-risk AI system is used in accordance with its intended purpose, and under conditions of reasonably foreseeable misuse (European Union, 2024, para. 2b).

Such a statement clearly points the finger at developers. Similarly, the chapter emphasizes the importance of risk management, transparency, bias mitigation, and human oversight. While Chapter 4 addresses transparency and comprises only Article 50, which once again focuses on deceptive practices, Chapter 7, on the other hand, focuses on Governance by establishing a governance framework comprising advisory forums, a scientific panel of independent experts, and defining national oversight. Once an overview of the Act itself is provided, it is possible to move to a critical analysis.

Risk, trust, and governance in the EU AI Act

The EU AI Act is undoubtedly a novelty in terms of harmonizing regulation across the European Union Member States. However, it is not immune to criticism. This section will identify three problematic elements of the Act, explain why they might be particularly challenging to interpret, and propose workarounds.

A rigid risk assessment model

A first critique might be moved against the apparent over-reliance on a fixed risk-based type of framework—prohibited AI practice (Chapter 2, Article 5), high-risk AI system (Chapter 3), limited risk (preamble, recital 53), and, reasonably, anything else outside these three categories as having minimal risk. Such a classification might overlook two types of issues. First, it might fail to capture the nuanced dimensions of all those AI models that do not fall within the Narrow AI (NAI) spectrum, that is, AI models not explicitly designed for a narrow (or Weak) specific set of tasks (e.g., language translation, face recognition, algorithm recommendation). General-purpose AI (GPAI) systems, therefore, even if arguably not yet present on the market (the commonly used ChatGPT for example, although versatile, might still be considered a type of NAI focused on NLP), could transition across all four different risk domains of the EU AI Act, which presents per se a challenging dilemma of how to classify and regulate such type of system. Second, NAI systems used for low-risk types of tasks, such as copywriting for a cooking-focused website, might be applied across different domains, including medical diagnosis. This is to say that AI systems are as dynamic as market needs; therefore, the narrow purpose of a specific type of AI model might change completely after deployment. This could lead to such a model being exploited for tasks other than those for which it was originally intended. Once again, AI risks are highly context-dependent, and market needs are, by nature, exploitable. It would be foolish to speculate that a specific model will be used exclusively for a specific task without having cross-functional implications for all the other models derived from its original model architecture.

The first thing to note is that the EU AI Act is not oblivious to the challenges posed by GPAI and addresses them directly (even though it does so only through models, not as independent systems). Recital 97 introduces GPAI models as components that can be integrated into AI

systems but remain distinct from these systems. This distinction means that GPAI models lack specific components that would make them viable for the public, as might be the case with end-user interfaces. Such a passage is fascinating, as it introduces the idea of risk inherent to GPAI, as proposed in recitals 110 and 111. These two recitals present the concept of systemic risk for GPAI models, recognizing that they can pose significant harm or disruption if not carefully supervised. For instance, High-impact capabilities (measured in floating-point operations, or FLOPs, and benchmarked against predefined thresholds) refer to the GPAI model's ability to perform tasks or influence outcomes at scale. The whole Chapter V, moreover, is entirely dedicated to GPAI and expands in detail several points concerning GPAI, such as the need for transparent technical documentation including training data (recital 107), the need for a code of practice (recital 116), as well as the previous two mentioned recitals on systemic risk and high-impact models (recitals 110 and 111).

The fact that GPAI is considered in the Act itself does not, however, mean that all possible struggles with it are accounted for. An important element note already mentioned in the previous sections—is that AI systems are as dynamic as market fluctuations, meaning they will necessarily evolve post-deployment. This means they have an adaptive nature. Such a situation might create contingencies that are not considered at the deployment stage, even if the system remains static during the initial period. A possible way around such a problem is periodic reassessment, ideally considering the public's response to the new release and recalibrating the safeguards accordingly. In this direction, Ceravolo and colleagues (Ceravolo, Damiani, D'Amico, et al., 2025) propose the HH4AI framework, a structured methodology for conducting human rights impact assessments under the EU AI Act, reinforcing the need to evaluate risk in multidimensional terms. Novelli and colleagues, for example, oppose to the rigid risk framework a dynamic-type risk assessment model, advocating for an *integrated model [that] enables the estimation of AI risk magnitude by considering the interaction between (a) risk determinants, (b) individual drivers of determinants, and (c) multiple risk types* (Novelli, Casolari, Rotolo et al., 2024, p. 2493) and "assessing AI risk through hazard chains, trade-offs among exposed values, vulnerability profiles, and cross sectorial risks provide a more accurate analysis of its risk. This approach turns the AIA risk categories into dynamic risk scenarios, changing with the interactions among factors, and ensures more proportionate regulatory measures" (Novelli, Casolari, Rotolo et al., 2024, p. 2495).

In addition to periodic reassessment, one might argue for continuous independent assessment of the AI system, validated by an independent institutional body. The Act might require built-in self-regulatory features that enable the system to automatically detect, flag, and report anomalies in its operation. Such flagging would then be submitted not only to the vendor's development team but also to the institutional body for independent review and corrective measures. Similarly, one might enhance public participation by introducing simplified, standardized AI systems and model cards for transparency and continuous monitoring. Such cards, which would ideally be easily retrievable and interpretable even for non-tech-savvy end-users, would summarize the system's capabilities, limitations, potential risks, and training data. Suppose a hospital implements an AI system to help prioritize ICU patients based on urgency. This system integrates a GPAI model trained to analyze patient vitals, symptoms, and medical history. To ensure transparency and safe usage, the hospital provides a system card to the medical staff—or easily mark the AI system with an identifier (unique ID) which can be accessed from any workstation on the hospital floor—detailing specific criteria such as purpose (priority given to specific risk scores), performance metrics (it validates system prioritization with human prioritization and gives out a percentage of alignment), limitations (it clarifies that certain conditions are under-represented in the training data hence when detected might skew the general assessment process), user guidance (how-to steps to override incorrect decision and amend historical data accordingly). Suppose now that a nurse notices that the system

consistently ranks elderly patients with specific symptoms lower on the priority list than expected. Referring to the system card, the nurse manually overrides the priority for one patient and files an anomaly report through the hospital's feedback system, which simultaneously reaches the AI service provider and the independent institutional body for review and corrective actions. Such an example can be easily replicated across many domains and instances.

The problem of trustworthiness

The EU AI Act makes trustworthy AI a central element, but how to achieve such trustworthiness might not be entirely clear. There appears to be a potential conflation between this concept and the idea of risk acceptability. Laux and colleagues (Laux, Wachter & Mittelstadt, 2024) argue, through an extensive analysis of the concept of trust and trustworthiness, that the EU AI Act appears to conflate the meaning of AI trustworthiness with mere technical compliance with the pre-established risk-based framework. Such framing reduces trustworthiness to a mechanical process of meeting regulatory thresholds—established by experts—rather than recognizing it as a more complex and ongoing socio-ethical process.

> Conflating trust and trustworthiness with the acceptability of risks blurs the distinction between acceptability judgments made by domain experts and the trustworthiness of AI systems implemented in society ... trustworthiness is a longitudinal concept that necessitates an iterative process of controls, communication, and accountability to establish and maintain its existence across both AI technologies and the institutions using them (Laux, Wachter & Mittelstadt, 2024, p. 4).

Therefore, such a conflation of trust and compliance risks eliminates public participation and an iterative accountability model based on end-user feedback, reducing requirements to mere checkboxes that stakeholders must check to get the go-ahead for development and market release. Consequently, public trust in AI models might be corrupted because the promised trustworthiness does not arise from public needs or feedback but rather from abstract guidelines designed by technocratic governance. On the same topic, Michael Gerlich (2023) provides a comprehensive analysis of perceptions and acceptance of AI across several European and non-European countries. As one might imagine, trust, perceived risks, societal attitudes, and cultural factors emerge as determinants of AI adoption, while risks and societal issues pose barriers. This is just to reiterate that trustworthy AI might be a concept far more complex than the elaboration in the EU AI Act. Therefore, such an idea risks being void; nothing more than a pro-forma statement printed on the page.

One might now argue that there is indeed a distinction between compliance and trust in the Act itself. A sign of this is found immediately in the initial recitals, such as recital 6, which states that "…as a prerequisite, AI should be a human-centric technology. It should serve as a tool for people, with the ultimate aim of increasing human well-being…" (European Union, 2024, recital 6). Articles 13 and 14, for example, ensure transparency, accountability, and human oversight, thereby building confidence in the system's reliability and safe operation. Similar references about the need to ensure or retain public trust can be found throughout the Act (see recital 59). It is necessary to remember also that the idea of trustworthiness is not something introduced ex-novo in the 2024 version of the Act, but rather the result of a much longer reflection inherited from the 2019 Ethics Guidelines for Trustworthy Artificial Intelligence compiled by High-Level Expert Group on AI(AI HLEG), which identified "…7 key requirements that AI systems should meet in order to be deemed trustworthy; Human agency and oversight, Technical Robustness and safety, Privacy and data governance, Transparency, Diversity, non-discrimination and fairness, Societal and environmental well-being and Accountability" (European Union, 2019). Such a list was subsequently finalized in the 2020 Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment

(European Union, 2020). This is to say that the idea of trustworthy AI is not new; rather, it has been a primary concern since the debate over AI's potential began. This, however, does not mean there is no room for improvement.

A possible approach might involve enhancing (already considered in Article 44) an Ethical Certification System designed by an independent task force (such as the AI HLEG of the EU) and tailored specifically for the evaluation of individual AI systems. Such a certification model should ideally be reviewed and updated every few years (in a manner similar to how many certifications are reviewed today) to ensure the AI system aligns with current needs and evolving public sensitivities. Furthermore, if an AI system were to fail such an assessment—ideally conducted at random by the independent task force—the AI service provider would be granted a set timeframe to address and rectify the identified shortcomings. Failure to do so would result in recurring penalties, calculated as a percentage of the system's revenue, as it is no longer compliant with EU regulations.

Suppose an AI service provider develops and deploys a high-risk AI system to assist in diagnosing rare diseases. Such systems are purchased and implemented by both government-operated and independently operated entities to support daily operations in the healthcare sector. The system is subject to an Ethical Certification System overseen by an independent task force. Every three years (or every five, as per Article 44), the task force evaluates the AI's compliance with ethical standards, focusing on transparency, accuracy, patient safety, and equitable treatment. During a routine, randomized audit, it is discovered that the system underperforms for certain ethnic groups, potentially leading to disparities in diagnostic outcomes. The task force issues a detailed report, and the AI service provider is given six months to address the issue by updating its training data and implementing corrective measures. If the company fails to comply within this time frame, it incurs escalating penalties, calculated as a percentage of the system's revenue.

There are many advantages to such a speculative assessment model, but three are particularly worth considering. The first is that randomized audits compel companies to stay current with national regulations and ethical standards, reducing the risk of non-compliance and maintaining trustworthiness. Moreover, linking penalties to a percentage of revenue generated by that specific AI system creates a strong financial incentive for immediate corrective measures, ensuring timely issue resolution. Finally, the whole certification process establishes a feedback loop, providing regular evaluations and recommendations that drive ongoing enhancements to the AI system's performance, safety, and fairness. Of course this is only a thought experiment, such a model could ideally be lifted and shifted to smaller scale AI systems, or ideally to any AI system directly interfacing with the public, or any AI system which bears any type of possible outcome for the general public (this is to say that the military sector, for example, could be excluded from such certification system).

Governance

There is a governance gap in the proposed EU AI Act model. Although the Act itself mandates conformity assessments (CAs) to ensure that high-risk AI systems meet the regulation's requirements (see, for example, Article 43), it provides limited practical guidance on how these assessments should be carried out, particularly regarding stakeholders' roles and responsibilities.

> Conformity assessments (CA) is a legal obligation, which must be fulfilled before a high-risk AI system is placed on the market. It aims at fostering accountability of AI providers. The EU AI Act defines a conformity assessment as the process of verifying whether the requirements set out in Title III, chapter 2 of the Regulation relating to an AI system have been fulfilled, while Title III offers provisions for high-risks systems only (Thelisson & Verma, 2024, p. 116).

Thelisson and Verma (Thelisson & Verma, 2024), in their analysis of the governance structure proposed by the EU AI Act for high-risk models, argue that while the Act aims to establish a robust regulatory framework, its provisions leave open questions about the practicalities of such CAs, particularly in terms of clarity, standardization, and enforcement mechanisms. For example, ex-ante controls include evaluations conducted before the deployment of high-risk AI systems, allowing both self-assessment by providers and external evaluation by notified bodies (see, for example, recitals 68, 125, and 126, and Articles 28 to 35).

> For high-risk AI systems listed in point 1 of Annex III, where, in demonstrating the compliance of a high-risk AI system with the requirements set out in Section 2, the provider has applied harmonised standards referred to in Article 40, or, where applicable, common specifications referred to in Article 41, the provider shall opt for one of the following conformity assessment procedures based on: (a) the internal control referred to in Annex VI; or (b) the assessment of the quality management system and the assessment of the technical documentation, with the involvement of a notified body, referred to in Annex VII. (European Union, 2024, art. 43).

The self-assessment element seems worrisome as it might create a conflict of interest for providers who clearly want to validate their product as compliant, and this despite the safeguard proposed by the EU AI Act (see, for example, ANNEXVI - Conformity assessment procedure based on internal control or ANNEXVII - Conformity based on an assessment of the quality management system and an assessment of the technical documentation). Yew and colleagues (Yew, Marino, Venkatasubramanian, 2025) highlight this very risk, introducing the concept of *avoision*—strategies by which providers may exploit loopholes to appear compliant while evading the substance of regulation.

Ex-post control, on the other hand, involves continuous monitoring and post-deployment compliance reviews, which presuppose a monitoring system. “post-market monitoring system’ means all activities carried out by providers of AI systems to collect and review experience gained from the use of AI systems they place on the market or put into service for the purpose of identifying any need to immediately apply any necessary corrective or preventive actions” (European Union, 2024, art. 3 para. 1).

As is the case with ex-ante processes, post-market ones also seem to lack certain pragmatic elements, despite multiple reiterations on their need throughout the Act (see, for example, Chapter IX - post-market monitoring, information sharing, and market surveillance). The apparent issue is that such a requirement seems too broad, thus leaving room for varied interpretation, especially concerning what constitutes the *proportionality* of the assessment (European Union, 2024, Article 72, para. 1). This could lead to inconsistent application across different providers, sectors, and AI systems. The first and most evident gap in such a self-assessment framework is the potential of providers to downplay risks or intentionally avoid classification as high-risk to bypass external scrutiny.

Once again, it is necessary to see how the Act addresses such potential gaps. The first thing to note is that the high-risk system classification is dynamic, meaning it can be modified to accommodate new technologies (Article 7). Similarly, Section 5 of Chapter III (Standards, Conformity Assessment, Certificates, Registration) establishes the framework for ensuring that high-risk AI systems meet EU safety and ethical standards. Finally, Chapter VI, particularly Articles 57 and 58, establishes the need to implement controlled environments with detailed guidelines (AI regulatory sandboxes) for regulatory compliance assessments and overall testing. Lewis and colleagues (Lewis, Lasek-Markey, Golpayegani et al., 2025) describe this adaptive approach as a *regulatory learning space*, emphasizing how sandboxes can serve as

sites of iterative governance and institutional learning under the Act.

A possible way to encourage compliance and avoid having stakeholders leverage the self-assessment gap to their advantage might be not only to implement sanctions but rather propose incentives such as tax breaks and grants to encourage providers to voluntarily adopt the highest standards; the classical carrot-and-stick approach. Stakeholders with a proven track record of consistently meeting EU AI regulations may request expedited CAs, which would ideally provide a fast-track assessment model, thus decreasing the market deployment time for new AI systems. This would, ipso facto, provide a market edge compared to potential competitors that were not as consistent in meeting EU regulations. From here, a wide range of initiatives could be implemented to attest to and publicize the strict adherence of such companies to EU regulations, such as an official EU Ethical Gold AI Label—similar to the Protected Designation of Origin quality scheme for agricultural products and foodstuffs.

Conclusions

The postulates upon which this paper is built are simple yet undeniable: AI is becoming increasingly integrated into the daily lives of many people and, as such, is shaping how the world is perceived and lived. Society is on the brink of a new technological revolution (if not already in the midst of it), which may enhance the integration between humans and technology, thereby pushing the human condition toward new transhuman horizons. It would be foolish, however, to turn a blind eye to the intrinsic dangers of such fast-paced changes: a potential blind faith in technological advancement for which proper moral and ethical safeguards are yet to be formulated. This paper did not propose addressing and solve such issues; instead, it aimed to provide a bird's-eye view of some of the most evident challenges with AI today. It proposed the idea that AI is a public matter. As such, it requires a social contract based on public participation and cooperation to achieve satisfactory results for AI regulation. Using Hobbesian principles as a framework, it was argued that there is a need for enforceable agreements that prioritize societal welfare over market-driven motives. The 2024 EU AI Act aligns nicely with such a narrative, as it represents the first comprehensive EU Regulation specifically devoted to artificial intelligence, with binding legal force across Member States. After providing a breakdown of some of the Act's most morally charged passages, the paper addressed its challenging key provisions, including the rigidity of the risk-based model and the focus on technical adherence and governance strategies for compliance assessment. A reinterpretation of these challenges was proposed by dwelling on the Act's own details before suggesting possible ways forward to enhance these provisions and addressing some of their most evident criticisms.

A final remark now must be made. The EU AI Act represents a monumental step forward in the ethical, moral, and legal domain of AI governance. Its success, however, depends on its adaptability, enforcement, and engagement. It was not the scope of the present paper to praise or criticize the Act; instead, the paper aimed to dissect some of its parts and place them in the broader AI discourse to better understand it. The Act is commendable and must be supported with intellectual and practical effort. Such an endeavor is not only a necessity but a moral duty; if AI truly possesses the potential to reshape human life as decisively as past technologies—think of the internet, for example—it is only right that everyone participates in its shaping. Today, we have both the means and the social awareness to grasp this potential, and ignorance is no excuse.

References

ALLOGHANI, M., AL-JUMEILY, D., MUSTAFINA, J. et al. (2020): *A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science*. In: M. W. Berry, A. Mohamed & Yap Bee Wah (ed.): *Supervised and unsupervised learning for data science*. Cham: Springer, pp. 3–23. https://doi.org/10.1007/978-3-030-22475-2.

AMERSHI, S., CAKMAK, M., KNOX, W. B. et al. (2014): Power to the people: The role of humans in interactive machine learning. In: *AI Magazine*, 35(4), pp. 105–120. https://doi.org/10.1609/aimag.v35i4.2513.
ARKOUDAS, K. & BRINGSJORD, S. (2014): *Philosophical foundations*. In: K. Frankish & W. M. Ramsey (ed.): *The Cambridge handbook of artificial intelligence*. Cambridge, UK: Cambridge University Press.
ASIMOV, I. (2014): *I, Robot*. Oxford: Macmillan.
BEAUCHAMP, T. L. & CHILDRESS, J. F. (2013): *Principles of biomedical ethics*. New York: Oxford University Press.
BECCARIA, C. (2018): *Dei delitti e delle pene: con una raccolta di lettere e documenti relativi alla nascita dell'opera e alla sua fortuna nell'Europa del Settecento*. Torino: Einaudi.
BERRY, D. (2023): The Limits of Computation: Joseph Weizenbaum and the ELIZA Chatbot. In: *Weizenbaum Journal of the Digital Society*, 3(3). https://doi.org/10.34669/WI.WJDS/3.3.2.
BISEN, V. S. (2022): *What is Human in the Loop Machine Learning: Why & How Used in AI?* https://medium.com/vsinghbisen/what-is-human-in-the-loop-machine-learning-why-how-used-in-ai-60c7b44eb2c0.
BUCHANAN, B. & SHORTLIFFE, E. (1985): *Rule-based expert systems: The mycin experiments of the stanford heuristic programming project. reprinted with corrections*. Reading, MA: Addison-Wesley.
BUTLER, S. (1987): *Erewhon*. Harmondsworth: Penguin.
BUTLIN, P., LONG, R., ELMOZNINO, E. et al. (2023): *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness*. In: arXiv. https://doi.org/10.48550/arXiv.2308.08708.
CARON, M. S. & GUPTA, A. (2020): *The Social Contract for AI*. In: arXiv. https://doi.org/10.48550/arXiv.2006.08140.
CERAVOLO, P., DAMIANI, E., D'AMICO, M. E. et al. (2025): *HH4AI: A methodological Framework for AI Human Rights impact assessment under the EUAI ACT*. In: arXiv. https://doi.org/10.48550/ARXIV.2503.18994.
CHANDLER, C., FOLTZ, P. W. & ELVEVÅG, B. (2022): Improving the applicability of AI for psychiatric applications through Human-in-the-loop methodologies. In: *Schizophrenia Bulletin*, 48(5), pp. 949–957. https://doi.org/10.1093/schbul/sbac038.
CHRISTIE, E. H., ERTAN, A., ADOMAITIS, L. et al. (2024): Regulating lethal autonomous weapon systems: exploring the challenges of explainability and traceability. In: *AI and Ethics*, 4(2), pp. 229–245. https://doi.org/10.1007/s43681-023-00261-0.
COHEN, A. P. (1993): Culture as identity: An anthropologist's view. In: *New Literary History*, 24(1), pp. 195–209. https://doi.org/10.2307/469278.
CORDESCHI, R. (2010): *Discovery of the artificial*. Dordrecht: Springer Netherlands.
DE FAUW, J., LEDSAM, J. R., ROMERA-PAREDES, B. et al. (2018): Clinically applicable deep learning for diagnosis and referral in retinal disease. In: *Nature Medicine*, 24(9), pp. 1342–1350. https://doi.org/10.1038/s41591-018-0107-6.
DILLON, S. (2020): The Eliza effect and its dangers: from demystification to gender critique. In: *Journal for Cultural Research*, 24(1), pp. 1–15. https://doi.org/10.1080/14797585.2020.1754642.
EUROPEAN UNION. (2017): *European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL))*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52017IP0051.
EUROPEAN UNION. (2019): *Ethics guidelines for trustworthy AI | Shaping Europe's digital future*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.
EUROPEAN UNION. (2020): *Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment | Shaping Europe's digital future*. [online], [Retrieved January 2, 2026].

Available at: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.
EUROPEAN UNION. (2024): *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance)*. https://eur-lex.europa.eu/eli/reg/2024/1689/oj.
FILIP, B. (2020): *The rise of neo-liberalism and the decline of freedom*. Cham: Springer International Publishing.
FLORIDI, L. (2019a):. What the near future of artificial intelligence could be. In: *Philosophy & Technology*, 32(1), pp. 1–15. https://doi.org/10.1007/s13347-019-00345-y.
FLORIDI, L. (2019b):. Translating principles into practices of digital ethics: Five risks of being unethical. In: *Philosophy & Technology*, 32(2), pp. 185–193. https://doi.org/10.1007/s13347-019-00354-x.
FLORIDI, L. & COWLS, J. (2019): A unified framework of five principles for AI in society. In: *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.8cd550d1.
FLORIDI, L., TADDEO, M. & TURILLI, M. (2009): Turing's Imitation game: Still an impossible challenge for all machines and some judges—An evaluation of the 2008 Loebner Contest. In: *Minds and Machines*, 19(1), pp. 145–150. https://doi.org/10.1007/s11023-008-9130-6.
FRANCÉS-GÓMEZ, P. (2023): Ethical principles and governance for AI. In: F. Lara & J. Deckers (eds.): *Ethics of artificial intelligence*. Cham: Springer, pp. 191–219.
GAUTHIER, D. P. (2006): *Morals by agreement*. Oxford: Oxford University Press.
GEISSLINGER, M., POSZLER, F., BETZ, J. et al. (2021): Autonomous driving ethics: from trolley problem to ethics of risk. In: *Philosophy & Technology*, 34(4), pp. 1033–1055. https://doi.org/10.1007/s13347-021-00449-4.
GERLICH, M. (2023): Perceptions and acceptance of Artificial Intelligence: A multi-dimensional study. In: *Social Sciences*, 12(9), pp. 1–24. https://doi.org/10.3390/socsci12090502.
GIGNAC, G. E. & SZODORAI, E. T. (2024): Defining intelligence: Bridging the gap between human and artificial perspectives. In: *Intelligence*, 104, 101832, pp. 1–16. https://doi.org/10.1016/j.intell.2024.101832.
GOTTFREDSON, L. S. (1997): Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. In: *Intelligence*, 24(1), pp. 13–23. https://doi.org/10.1016/S0160-2896(97)90011-8.
GREENE, J. D., NYSTROM, L. E., ENGELL, A. D. et al. (2004): The neural bases of cognitive conflict and control in moral judgment. In: *Neuron*, 44(2), pp. 389–400. https://doi.org/10.1016/j.neuron.2004.09.027.
HAGENDORFF, T. (2020): The ethics of AI ethics: An evaluation of guidelines. In: *Minds and Machines*, 30(1), pp. 99–120. https://doi.org/10.1007/s11023-020-09517-8.
HASSAN, N. M., HAMAD, S. & MAHAR, K. (2022): Mammogram breast cancer CAD systems for mass detection and classification: a review. In: *Multimedia Tools and Applications*, 81(14), pp. 20043–20075. https://doi.org/10.1007/s11042-022-12332-1.
HELION, C. & PIZARRO, D. A. (2015): *Beyond dual-processes: The interplay of reason and emotion in moral judgment*. In: J. Clausen & N. Levy (eds.): *Handbook of Neuroethics*. Dordrecht: Springer Netherlands, pp. 109–125. https://doi.org/10.1007/978-94-007-4707-4_160.
HERRERA, F. & CALDERÓN, R. (2025): *Opacity as a feature, not a flaw: The LoBOX governance ethic for role-sensitive explainability and institutional trust in AI*. In: arXiv.

https://doi.org/10.48550/ARXIV.2505.20304.
HO, K. (2005): Situating global capitalisms: A view from Wall Street investment banks. In: *Cultural Anthropology*, 20(1), pp. 68–96.
HOBBES, T. (1997): *Leviathan: or the matter, forme, & power of a common-wealth ecclesiastical and civill*. New York: Atria Books.
JALŠENJAK, B. (2020): The Artificial Intelligence singularity: What it is and what it is not. In: S. Skansi (ed.): *Guide to Deep Learning Basics*. Cham: Springer International Publishing, pp. 107–115. https://doi.org/10.1007/978-3-030-37591-1_10.
JENKINS, R., ČERNÝ, D. & HŘÍBEK, T. (2022): *Autonomous vehicle ethics: The trolley problem and beyond*. New York, NY: Oxford University Press. https://doi.org/10.1093/oso/9780197639191.001.0001.
JONES, C. R. & BERGEN, B. K. (2023): *Does GPT-4 pass the Turing test?* In: arXiv. https://doi.org/10.48550/ARXIV.2310.20216.
JONES, C. R. & BERGEN, B. K. (2024): *People cannot distinguish GPT-4 from a human in a Turing test*. In: arXiv. https://doi.org/10.48550/ARXIV.2405.08007.
JULIANI, A., ARULKUMARAN, K., SASAI, S. et al. (2022): *On the link between conscious function and general intelligence in humans and machines*. In: arXiv. https://doi.org/10.48550/arXiv.2204.05133.
KEELING, G. (2020): Why trolley problems matter for the ethics of automated vehicles. In: *Science and Engineering Ethics*, 26(1), pp. 293–307. https://doi.org/10.1007/s11948-019-00096-1.
KLINE, R. (2011): Cybernetics, Automata Studies, and the Dartmouth Conference on Artificial Intelligence. In: *IEEE Annals of the History of Computing*, 33(4), pp. 5–16. https://doi.org/10.1109/MAHC.2010.44.
LAUX, J., WACHTER, S. & MITTELSTADT, B. (2024): Trustworthy artificial intelligence and the European Union AI Act: On the conflation of trustworthiness and acceptability of risk. In: *Regulation & Governance*, 18(1), pp. 3–32. https://doi.org/10.1111/rego.12512.
LEAVY, S., O'SULLIVAN, B. & SIAPERA, E. (2020): *Data, power and bias in Artificial Intelligence*. In: arXiv. https://doi.org/10.48550/ARXIV.2008.07341.
LEWIS, D., LASEK-MARKEY, M., GOLPAYEGANI, D. et al. (2025): *Mapping the Regulatory Learning Space for the EU AI Act*. In: arXiv. https://doi.org/10.48550/ARXIV.2503.05787.
LIN, Z. (2025): Beyond principlism: Practical strategies for ethical AI use in research practices. In: *AI and Ethics*, 5, pp. 2719–2731. https://doi.org/10.1007/s43681-024-00585-5.
LOCKE, J. (2010): *Two treatises of government: in the former, the false principles and foundation of Sir Robert Filmer, and his followers are detected and overthrown; the latter is an essay concerning the true original, extent, and end of civil-government*. Clark, NJ: Lawbook Exchange, LTD.
LOPES, D. (2023): *Towards the artificial brain: A base framework for modelling consciousness and unconsciousness*. In: arXiv. https://doi.org/10.48550/arXiv.2305.08863.
MACFARLANE, B. (1999): Re-evaluating the Realist Conception of War as a Business Metaphor. In: *Teaching Business Ethics*, 3(1), pp. 27–35.
https://doi.org/10.1023/A:1009753807317.
MALLIORI, A. & PALLIKARAKIS, N. (2022): Breast cancer detection using machine learning in digital mammography and breast tomosynthesis: A systematic review. In: *Health and Technology*, 12(5), pp. 893–910. https://doi.org/10.1007/s12553-022-00693-4.
MARMOLEJO-RAMOS, F., WORKMAN, T., WALKER, C. et al. (2022): AI-powered narrative building for facilitating public participation and engagement. In: *Discover Artificial Intelligence*, 2(1), p. 7. https://doi.org/10.1007/s44163-022-00023-7.
MCNAMARA, A., SMITH, J. & MURPHY-HILL, E. (2018): *Does ACM's code of ethics*

change ethical decision making in software development? Lake Buena Vista FL: ACM. https://doi.org/10.1145/3236024.3264833.
MENTXAKA, O., DÍAZ-RODRÍGUEZ, N., COECKELBERGH, M. et al. (2025): *Aligning trustworthy AI with democracy: A dual taxonomy of opportunities and risks*. In: arXiv. https://doi.org/10.48550/ARXIV.2505.13565.
MESTVIRISHVILI, M., MESTVIRISHVILI, N., KVITSIANI, M. et al. (2020): Emotional intelligence for moral character: Do emotion-related competencies lead to better moral functioning? In: *Psychological Studies*, 65(3), pp. 307–317. https://doi.org/10.1007/s12646-020-00564-w.
MONTRÉAL DECLARATION ON RESPONSIBLE AI (2018a): [online], [Retrieved February 10, 2026]. Available at: https://montrealdeclaration-responsibleai.com/.
MOOR, J. (2006): The Dartmouth College Artificial Intelligence conference: The next fifty years. In: *AI Magazine*, 27(4), pp. 87–91.
MOSQUEIRA-REY, E., HERNÁNDEZ-PEREIRA, E., ALONSO-RÍOS, D. et al. (2023): Human-in-the-loop machine learning: a state of the art. In: *Artificial Intelligence Review*, 56(4), pp. 3005–3054. https://doi.org/10.1007/s10462-022-10246-w.
NATALE, S. (2021): *Deceitful media*: *Artificial Intelligence and social life after the Turing test*. New York: Oxford University Press. https://doi.org/10.1093/oso/9780190080365.003.0004.
NOVELLI, C., CASOLARI, F., ROTOLO, A. et al. (2024): Taking AI risks seriously: a new assessment model for the AI Act. In: *AI & Society*, 39(5), pp. 2493–2497. https://doi.org/10.1007/s00146-023-01723-z.
ONEŢ, A.-E. & CIOCOI-POP, A.-B. (2022): Of battle and business: Military language in the corporate environment. In: *International conference KNOWLEDGE-BASED ORGANIZATION*, 28(2), pp. 213–218. https://doi.org/10.2478/kbo-2022-0075.
PASQUALE, F. (2015): *The black box society: the secret algorithms that control money and information*. Cambridge, MA: Harvard University Press.
PAULO, N. (2023): The trolley problem in the ethics of autonomous vehicles. In: *The Philosophical Quarterly*, 73(4), pp. 1046–1066. https://doi.org/10.1093/pq/pqad051.
PAULY, P. J. (1982): Samuel Butler and his Darwinian critics. In: *Victorian Studies*, 25(2), pp. 161–180.
POWLES, J. & HODSON, H. (2017): Google DeepMind and healthcare in an age of algorithms. In: *Health and Technology*, 7(4), pp. 351–367. https://doi.org/10.1007/s12553-017-0179-1.
RAHWAN, I. (2018): Society-in-the-loop: programming the algorithmic social contract. In: *Ethics and Information Technology*, 20(1), pp. 5–14. https://doi.org/10.1007/s10676-017-9430-8.
RATHI, I., TAYLOR, S., BERGEN, B. K. et al. (2024): *GPT-4 is judged more human than humans in displaced and inverted Turing tests*. In: arXiv. https://doi.org/10.48550/arXiv.2407.08853.
RAWLS, J. (2005): *A theory of justice: Original edition*. Cambridge, MA: Harvard University Press. https://doi.org/10.2307/j.ctvjf9z6v.
Robotics Openletter to the European Commission. (2018b):. https://robotics-openletter.eu/.
ROSS, P. & MAYNARD, K. (2021): Towards a 4th industrial revolution. In: *Intelligent Buildings International*, 13(3), pp. 159–161. https://doi.org/10.1080/17508975.2021.1873625.
ROUSSEAU, J.-J. (2012): *On the social contract*. Newburyport: Dover Publications.
SANDU, A. & NISTOR, P. (2021): Digital Dementia. In: *Eastern-European Journal of Medical Humanities and Bioethics*, 4(1), pp. 1–6. https://doi.org/10.18662/eejmhb/4.1/22.
SCHWAB, K. (2016): *The Fourth Industrial Revolution: what it means and how to respond*. https://www.weforum.org/stories/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/.

SETH, J. (2024): *Public Perception of AI: Sentiment and Opportunity*. In: arXiv. https://doi.org/10.48550/arXiv.2407.15998.
SHARMA, V., GOYAL, M. & MALIK, D. (2017): An intelligent behavior shown by Chatbot system. In: *International Journal of New Technology and Research*, 3(4), p. 263312.
SHRAGER, J. (2024): *ELIZA Reinterpreted: The world's first chatbot was not intended as a chatbot at all*. In: arXiv. https://doi.org/10.48550/ARXIV.2406.17650.
SKYRMS, B. (2014): *Evolution of the social contract*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139924825.
SPITZER, M. (2012): *Digitale Demenz: wie wir uns und unsere Kinder um den Verstand bringen* [*Digital dementia: How we are driving ourselves and our children crazy*]. München: Droemer.
THEA VON HARBOU (2015): *Metropolis*. S.l.: Edcon Publishing.
THELISSON, E. & VERMA, H. (2024): Conformity assessment under the EU AI act general approach. In: *AI and Ethics*, 4(1), pp. 113–121. https://doi.org/10.1007/s43681-023-00402-5.
TUPASELA, A. & DI NUCCI, E. (2020): Concordance as evidence in the Watson for Oncology decision-support system. In: *AI & Society*, 35(4), pp. 811–818. https://doi.org/10.1007/s00146-020-00945-9.
VON ESCHENBACH, W. J. (2021): Transparency and the Black Box Problem: Why we do not trust AI. In: *Philosophy & Technology*, 34(4), pp. 1607–1622. https://doi.org/10.1007/s13347-021-00477-0.
WALKER, R., DILLARD-WRIGHT, J. & IRADUKUNDA, F. (2023): Algorithmic bias in artificial intelligence is a problem—And the root issue is power. In: *Nursing Outlook*, 71(5), p. 102023. https://doi.org/10.1016/j.outlook.2023.102023.
WANG, L. (2024): Mammography with deep learning for breast cancer detection. In: *Frontiers in Oncology*, 14, 1281922, pp. 1–16. https://doi.org/10.3389/fonc.2024.1281922.
WANG, P. (2019): On defining Artificial Intelligence. In: *Journal of Artificial General Intelligence*, 10(2), pp. 1–37. https://doi.org/10.2478/jagi-2019-0002.
WEIZENBAUM, J. (1966): ELIZA—a computer program for the study of natural language communication between man and machine. In: *Communications of the ACM*, 9(1), pp. 36–45. https://doi.org/10.1145/365153.365168.
YEW, R.-J., MARINO, B. & VENKATASUBRAMANIAN, S. (2025): *Red Teaming AI Policy: A Taxonomy of Avoision and the EU AI Act*. In: arXiv. https://doi.org/10.48550/ARXIV.2506.01931.
YIGITCANLAR, T., DEGIRMENCI, K. & INKINEN, T. (2024): Drivers behind the public perception of artificial intelligence: insights from major Australian cities. In: *AI & SOCIETY*, 39(3), pp. 833–853. https://doi.org/10.1007/s00146-022-01566-0.
ZHAN, H. & WAN, D. (2024): Ethical considerations of the trolley problem in autonomous driving: A philosophical and technological analysis. In: *World Electric Vehicle Journal*, 15(9), 404, pp. 1–11. https://doi.org/10.3390/wevj15090404.