

Artificial Intelligence and Humanity: the urgency for an ethical and conscious coexistence

Inês Gomes Miranda¹

Judite Gonçalves de Freitas²

Abstract: The present study proposes a critical and constructive reflection on the role of Artificial Intelligence (AI) in contemporary societies, focusing on the ethical, political, legal, and social challenges it raises, its possible repercussions for democracy and on the visions of the future it evokes. Far from advocating a technophobic stance or a rejection of AI, this work proceeds from the conviction that the conscious, responsible, and ethical integration of these technologies into human life is not merely necessary but indeed inevitable – much like what is defended in relation to the coexistence between human beings and nature: a coexistence guided by values of respect, balance, responsibility, and interconnection. Through a theoretical and bibliographical analysis, the study begins with an introduction to the evolution of AI and its most significant applications, seeking to demonstrate its revolutionary potential. It then explores the complexity of the ethical dilemmas arising from its use, such as algorithmic bias and democracy, accountability, and the instrumentalization of the human being, while also undertaking a brief incursion into the political, legal, and social dilemmas involved. In a third stage, the analysis turns to the relationship between AI and democracy, with particular emphasis on the risks associated with information manipulation, digital surveillance, the concentration of technological power, and the erosion of public debate. Finally, the study reflects on the urgency of investing in digital literacy, the central role of education in preserving intellectual autonomy and the fragile balance between security and freedom.

Keywords: Artificial Intelligence; Ethics; Algorithms; Democracy; Accountability; Digital Literacy.

Introduction

Over recent decades, Artificial Intelligence (AI) has consolidated its position as one of the most potentially transformative forces of contemporary society. From facial recognition systems to the automation of decision-

¹ Aluna da Universidade Fernando Pessoa/Porto/Portugal.

ORCID: <https://orcid.org/0009-0007-5266-0341>. Email: igm.profissional.2@gmail.com

² Universidade Fernando Pessoa/Faculdade de Ciências Humanas e Sociais/Porto/Portugal.

CEPESE e IRIP/UNova/Portugal.

ORCID: <https://orcid.org/0000-0003-4516-9988>. Email: jfreitas@ufp.edu.pt



-making processes across public and private sectors, its influence, both on the pace and on the modalities through which individuals live, work, and relate to one another, has grown markedly. This evolution, grounded in scientific and technological development, is inseparable from the major social and political questions of the twenty-first century. Against this backdrop, the present study advances a critical and constructive examination of the integration of AI, premised on the conviction that its presence in human societies must be shaped by ethical, democratic, and humanistic values.

Rather than advocating a technophobic stance or a rejection of technological progress, a position that would be both implausible and analytically unproductive, this work argues for the need to conceptualize AI in a manner analogous to an idealized relationship between human beings and nature: a coexistence oriented by responsibility, mutual respect, and the pursuit of equilibrium, not by logics of domination. Accordingly, AI is not treated as an autonomous and inherently hostile force, but, when accompanied by adequate capacity-building and critical empowerment of individuals, as an instrument capable of directing technological progress toward the reinforcement of collective life grounded in ethical and democratic principles.

Drawing on a theoretical framework and an extensive literature review, the study is structured into four chapters. The first offers an overview of the evolution of AI and its principal applications, which is essential to apprehend both the breadth of its technological impact and its transformative potential. The second examines the ethical, political, legal, and social challenges associated with AI, including algorithmic bias, the dehumanization of decision-making, and the need for regulatory harmonization, underscoring the urgency of strategies oriented toward transparency and accountability. The third chapter analyses the intersections between AI and democratic regimes, highlighting the risks of informational manipulation, surveillance, and concentration of power, as well as their implications for the erosion of fundamental freedoms and democratic principles. The fourth and final chapter reflects on the urgency of digital literacy for full citizenship, the role of education in fostering intellectual autonomy, and the challenge of balancing security and freedom in the context of ongoing technological advancement.

By advancing this reflection, the study seeks to contribute to an increasingly pressing debate: how can AI be integrated in such a way as to preserve, and potentially even deepen, what is understood as human? The answer, necessarily provisional, also requires recognizing that the future is not pre-programmed; it will depend on the collective choices made in the present.

1. Artificial Intelligence: Definition and Evolution

Artificial intelligence (AI) has become one of the most influential and disruptive fields in contemporary science and technology, securing a central place in academic, political, and social debate over recent decades. As a genuine technological upheaval, stemming from its rapid evolution and profound impact on the organisation of societies (Tallberg *et al.*, 2023), it imposes itself as a privileged object of critical inquiry.

This chapter therefore begins with a brief historical contextualisation, highlighting key milestones in AI's development, from early logical algorithms to the revolution brought about by Deep Learning (DL) and Generative Models that characterise the present moment. Adopting a diachronic perspective, it establishes a conceptual foundation for AI, encompassing classical definitions centred on information processing as well as more recent approaches grounded in machine learning and data analysis. The chapter seeks to address a guiding question: what is meant by AI, and which dimensions does it entail?

1.1. Artificial Intelligence: stages of development and transformation

What is AI? In concise terms, it may be understood as the attempt to simulate, through computational systems, capacities that correspond to, or may eventually surpass, human intelligence. Although multiple definitions coexist, a widely cited formulation is that of the American mathematician Richard Bellman (1978), who characterises AI as the automation of activities associated with human thinking, such as decision-making, problem solving, and learning.

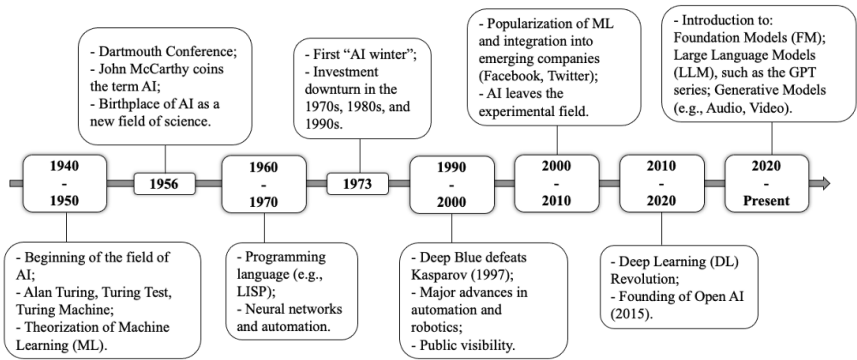
This definition raises a second question: what is intelligence? Intelligence is a multifaceted concept and, much like AI itself, remains subject to diverse interpretations (Russell & Norvig, 2021). According to the American Psychological Association (APA), intelligence can be described as the capacity to infer information, learn from experience, adapt to the environment, and understand and appropriately employ thought and reason (APA 2018). Yet what, then, still distinguishes human intelligence from AI? Even though AI is capable of performing highly complex tasks, there remains something singular about the human mind.

A third question thus emerges: that of consciousness. Consciousness may be defined as the capacity to experience, in an aware manner, the self and the surrounding environment (Stanford Encyclopedia of Philosophy, 2014). Consciousness – understood as the subjective experience of the world – constitutes a key feature separating human intelligence from AI. This capacity, which combines an understanding of the world with self-awareness, has not yet been attributed to AI systems and remains without a definitive explanation, generating both concern and enthusiasm (Baerren, 2025). For Damásio (2000), consciousness is an entirely private, first-person phenomenon that unfolds within the mind, and this understanding makes it possible, at least from a current philosophical and scientific standpoint, to draw a line between AI and biological intelligence.

Although the roots of AI can be traced back to Ancient Greece, particularly to Aristotle (384-322 BCE) and the introduction of deductive reasoning, which provides conceptual foundations for this technology (Rai, 2024), the analysis developed here begins with the contribution of Alan Turing, mathematician and computer scientist, as illustrated in Figure 1.

In 1957, Bellman formalised a class of sequential decision problems known as Markov Decision Processes (MDPs), which are fundamental for understanding decision-making in uncertain and dynamic environments and are now widely applied in robotics and automation.

Figure 1
Diachronic trajectory of AI evolution



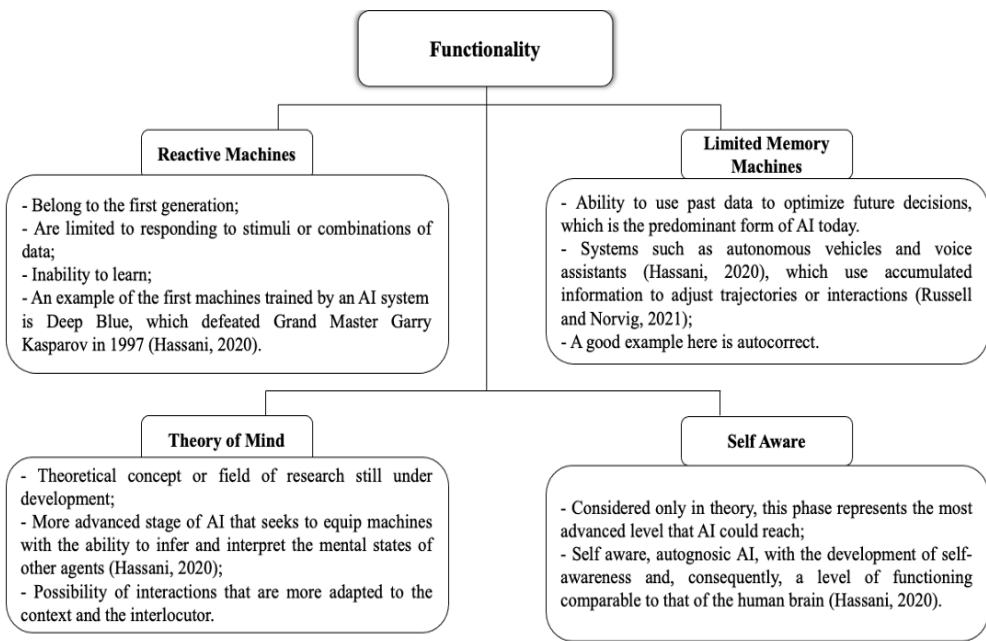
Source: Adapted from Daylight (2015), Smith *et al.* (2006), and Rai (2024).

1.2. Classification of AI: Functionality and Capacity

According to Hassani *et al.* (2020), AI may be defined as a computational system capable of reproducing human cognitive processes, drawing on data from multiple sources in order to make decisions and learn from patterns. Like all systems and machines built upon it, AI can be classified in several ways, including (but not limited to) functionality, domains of application, and level of autonomy. In this subsection, two distinct classification dimensions are examined, functionality and capacity, as proposed by Hassani.

The first refers to the “mind” of AI, that is, its ability to think with greater or lesser autonomy and the extent to which it resembles human thought. Within this dimension, four subcategories can be identified: Reactive Machines, Limited Memory Machines, Theory of Mind, and self-aware AI, represented in Figure 2.

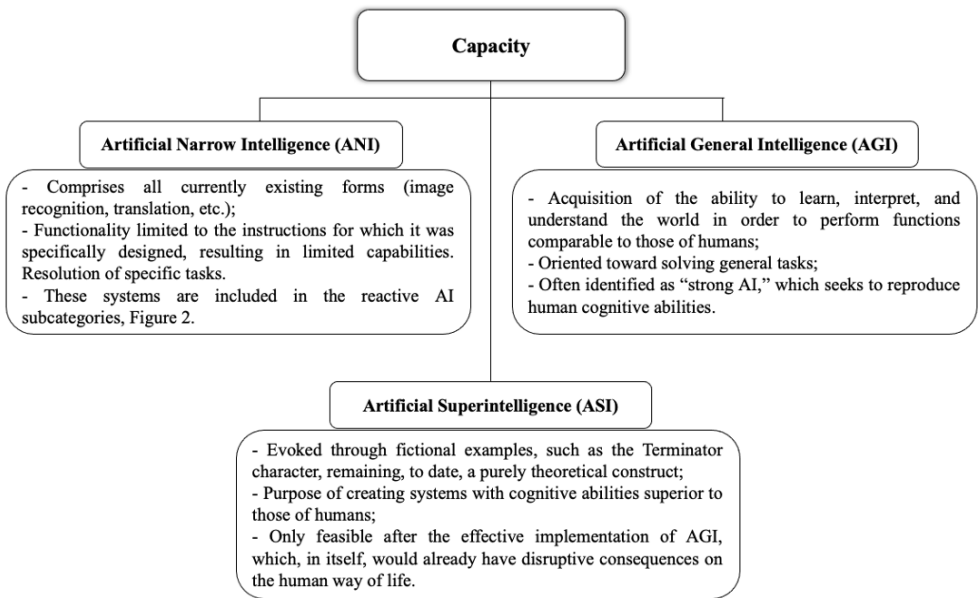
Figure 2
Classification of AI by functionality



Source: Adapted from Hassani *et al.* (2020).

The second classification category is grounded in technological capacity and distinguishes three subtypes: Artificial Narrow Intelligence (ANI), Artificial General Intelligence (AGI), and Artificial Superintelligence (ASI), shown in Figure 3.

Figure 3
Classification of AI by capacity



Source: Adapted from Hassani *et al.* (2020).

1.3. Technological advances and language processing models

According to Jeff Crume (2024), before the consolidation of Machine Learning (ML), AI relied predominantly on traditional programming approaches. ML, which gained prominence from the 2010s onwards, marked a turning point by enabling systems to learn from large datasets and prove particularly effective in prediction and pattern-detection tasks, with relevant applications in cybersecurity and information analysis. Deep Learning (DL), developed in the same period and based on deep neural networks inspired by the functioning of the human brain, exponentially expanded processing and recognition capabilities, supporting substantial progress in the simulation and mimicry of cognitive functions.

The 2020s witnessed the consolidation of Generative AI (GenAI), now embedded in widely deployed applications such as recommendation systems (e.g. Netflix, Spotify) and virtual assistants (e.g. Alexa, Siri). In 2022, the public release of ChatGPT – built on OpenAI’s GPT model family, initiated in 2015 and iteratively developed since 2018 – represented a decisive milestone in the popularisation of these systems. Grounded in Foundation Models (FM) trained on massive volumes of data and still situated within the scope of Artificial Narrow Intelligence (ANI), these systems inaugurated a new stage in AI by placing Large Language Models (LLM) at the centre of the debate, distinguished by their capacity to engage in dialogue, translate, summarise and automatically generate complex texts. Their diffusion not only democratised access to advanced tools and turned

them into a global cultural phenomenon, but also fuelled the emergence of multimodal applications capable of generating images, audio and video, giving rise to phenomena such as deepfakes.

The reach of these technologies goes far beyond strictly academic domains, as they are now embedded in writing tools, search engines, educational platforms, entertainment applications and decision-support systems (Rashid & Kausik, 2024). This ubiquity helps explain why LLMs are often considered emblematic of the current phase of AI, while simultaneously fuelling debates on creativity, authenticity and accountability. Although these developments confirm the transformative nature of the technology, they also raise a series of ethical, political, legal and social challenges, which will be examined in the following section.

2. Ethical, political, legal and social challenges of AI

Alongside being a driver of efficiency and innovation, AI has the potential to significantly reshape social dynamics, reconfiguring decision-making processes, power structures and prevailing understandings of human dignity. For these reasons, the scientific community is currently engaged in intense debate on the ethical, political and legal regulation of AI, as well as on its broader social impacts.

This chapter synthesises the main emerging challenges and some of the corresponding responses, taking into account the context of growing interdependence and interaction between humans and AI. To this end, the discussion focuses on three main axes: the ethical void associated with AI, the crisis of responsibility arising from algorithmic decision-making, and the obstacles that hinder the construction of regulatory frameworks capable of addressing the challenges at stake.

2.1. Beyond the ethical void: from uncertainty to trust

As the replacement of human labour by machines in the name of resource optimisation becomes a reality, questions arise regarding the manner in which human–technology interactions are being configured (Rashid & Kausik, 2024). As Rashid and Kausik emphasise, it will be necessary to find a balance between innovation and ethical principles in order to ensure that this relationship maximises benefits while minimising risks.

In July 2020, the High-Level Expert Group on AI (AI HLEG) presented to the European Commission the Assessment List for Trustworthy AI (ALTAI), the final version of a set of recommendations for AI development. According to this document, in addition to complying with existing legal frameworks, an AI system can be considered trustworthy if it adheres to seven ethical principles, represented in Figure 4.

Figure 4
Seven ethical principles for trustworthy AI

1) Agency and human oversight	AI should promote human autonomy, empowering individuals to make informed decisions. At the same time, oversight mechanisms should be ensured that follow human-in-the-loop, human-on-the-loop, and human-in-command approaches.
2) Technical robustness and safety	AI systems must be secure and resilient, with backup plans capable of handling errors and minimizing damage. They must be accurate, reliable, and reproducible.
3) Privacy and data governance	In addition to ensuring respect for privacy and personal data, data management mechanisms should be put in place to ensure the quality and integrity of information, while guaranteeing legitimate access and protection against abusive, discriminatory, or harmful uses.
4) Transparency	AI systems, data quality, models, and processes must be traceable through mechanisms designed for this purpose. Everyone who interacts with AI must be aware of its capabilities and limitations.
5) Equity, diversity, and non-discrimination	Access to AI systems must be guaranteed for all, avoiding inequalities and discrimination.
6) Social and environmental well-being	Everyone should be able to enjoy the benefits that AI brings, including future generations and the environment. Ecological responsibility should be promoted, considering the social and environmental impacts it entails.
7) Accountability	The use of AI requires the existence of algorithmic auditing mechanisms that enable accountability for its results and effects. Remedies and compensation must also be provided for.

Source: Adapted from Floridi (2019) and the European Commission (2019).

According to Luciano Floridi, a member of the HLEG, ethical uncertainty generates both unnecessary risks and excessive caution, preventing society from fully benefiting from the technology, whether through underuse, disproportionate use or malicious appropriation (Floridi, 2019). The guidelines seek to steer AI development on the basis of political and social consensus, involving broad deliberation and a long-term view of its purposes and priorities. Despite their specifically European origin and their integration within a wider landscape of ethical frameworks, this value system aspires to global reach. To succeed, however, it will have to overcome a range of normative, geopolitical and social challenges (Floridi, 2019 and 2022). The following pages address these challenges in detail.

2.2. Algorithmic decision-making: autonomy and the crisis of responsibility

As decision-making processes are increasingly delegated to AI systems, new debates emerge regarding where responsibility ultimately lies. From autonomous vehicles – capable of taking decisions within fractions of a second, with direct, immediate and potentially irreversible consequences for human life – to algorithms used in medical diagnosis, judicial decision-making or recruitment processes, it becomes ever more urgent to determine who is accountable for the errors, harms or injustices resulting from such choices (Han, 2015).

Situated within the so-called Fourth Industrial Revolution, characterised by the centrality of automation and the replacement of human tasks with more efficient processes, AI systems herald profound transformations across multiple sectors (Rashid & Kausik, 2024). From healthcare to the military, and including telecommunications, finance, education, manufacturing, agriculture, energy and the automotive industry, AI has driven an exponential increase in productive capacity and accelerated decision-making processes,

including those in the judicial domain (Vasdani, 2020). However, this progress is accompanied by a critical problem of interpretability: systems make decisions but cannot explain in a transparent manner the reasons underlying those decisions, which should in itself – and especially from a legal standpoint – render such decisions unacceptable (Benjamins & Salazar García, 2019).

Moreover, AI currently operates only in an apparently autonomous fashion: “it depends on human input for data collection, classification and maintenance and, although it can identify patterns, it remains unable to grasp their meaning, as in the case of cancer cells” (Dignum 2023, p. 199). This reinforces the argument for the need to assign responsibility for the decisions produced. Regardless of the ongoing debate on whether a “robot legal personality” should be created (Benjamins & Salazar García, 2019, p. 92), Virginia Dignum (2023) emphasises that AI systems are artefacts created by human beings and that ultimate responsibility for their use must rest with those who design and deploy them.

2.3. Three Key Challenges in AI Regulation

Recognizing that the debate surrounding AI regulation spans multiple dimensions that cannot be exhaustively covered here, this study prioritizes a synthetic overview of three structural legal and political challenges that constrain its implementation.

The first challenge is the so-called pacing problem (Yordanova & Bertels, 2024). The interval between the proposal, voting, and enactment of legislation often allows technology to outpace the law, creating a near-perpetual state of legal vacuum conducive to unregulated development and inadequately monitored application and use. This asymmetry highlights not only the disparity between the lifecycle of AI technologies and political or human cycles, but also underscores the necessity for political actors to ensure regulatory continuity and oversight regardless of shifts in political power.

The second challenge concerns the problem of taxonomy, emphasized by the HLEG (Yordanova & Bertels, 2024): the absence of a single, consensual definition of AI. The plurality of concepts and definitions generates ambiguity regarding the object of regulation, compromising normative clarity and effectiveness. As Krafft *et al.* (2020) point out, this conceptual fragmentation complicates policy formulation, since researchers tend to adopt functional definitions focused on the technical capabilities of current technologies, whereas policymakers favour comparisons with human cognition and behaviour projected onto hypothetical scenarios. This divergence can lead to an overemphasis on future risks at the expense of addressing current challenges (Krafft *et al.*, 2020).

The third challenge relates to the transnational nature and scale of AI, which contrasts sharply with the absence of centralized global governance (Tallberg *et al.*, 2023). Despite legal advances promoted by the EU (AI Act, GDPR), the US (National AI Initiative Act), and China (AI Ethics Code) – among others – and private sector efforts towards self-regulation (Meta’s GenAI Content Risk Standards), the current regulatory framework proves insufficient to manage the cross-border externalities of AI (Tallberg *et al.*, 2023, p. 9). Moreover, this multiplicity of initiatives contributes to an increasing “mushrooming of declarations” (Floridi, 2019), producing inconsistencies and confusion among stakeholders. This process results in the creation of a “supermarket of principles” (Floridi, 2019, p. 2), a system of selective preferences and pick-and-choose that hampers adaptation to a common ethical-normative framework (Floridi, 2019).

3. Democracy, Algorithms, and Artificial Intelligence

The obstacles identified in the preceding chapter are part of a broader debate on the impact of AI on political institutions and democratic quality. Regulating AI involves not only technical and legal dimensions but also considerations regarding the role of information in public deliberation, opinion formation, and democratic legitimacy.

This chapter focuses on the intersection of technology and democracy, exploring dynamics of information, disinformation, and manipulation, the effects of filter bubbles (Pariser, 2011), and the growing influence of algorithmic systems on the public sphere. The aim is to understand the challenges posed to democratic plurality, reflect on possible transformations in democratic judgment, and consider the concept of algocracies as identified by some authors.

3.1. (Dis)Information, Filter Bubbles, and Opinion Manipulation

Algorithmic communication – the management of information through systems that select, prioritize, and personalize content – presents serious challenges to democracy. By facilitating the diffusion of (dis)information, it has immense potential to corrode the shared knowledge base that underpins consensus and democratic institutions (Christiano, 2022). As Harari (2024) observes, the core issue at hand is primarily a network problem, more specifically an information problem. The distortion of information on which human societies depend inevitably compromises the quality of our decisions, regardless of intelligence levels.

In 2011, Eli Pariser warned of the algorithmic logic of platforms like Google and Facebook, now common across major platforms, where content personalization delivers the information users want rather than need, contributing to unprecedented intellectual isolation. This leads to the formation of what Pariser terms filter bubbles: information spheres that restrict access to diverse information, reinforce pre-existing tendencies, and isolate individuals from the community (Pariser, 2011). This process progressively erodes the common cognitive heritage that supports social consensus and democratic institutions.

While algorithmic manipulation is complex, Thomas Christiano (2022) identifies three processes manifesting in the public sphere: filtering, hypernudging, and microtargeting. Filtering involves the selection and presentation of information to users, limiting access to alternative content and reinforcing filter bubbles. Hypernudging refers to the continuous reconfiguration of accessible content, influencing choice patterns without users' awareness. Microtargeting involves directing messages to specific segments by exploiting personal data and vulnerabilities to enhance persuasive effects.

As the digital environment matures, social networks and digital platforms have consolidated not only as channels for socialization, commerce, and entertainment but also as tools for ideological dissemination and political mobilization (e.g., the Cambridge Analytica case). Although these processes need not be manipulative or harmful, the algorithmic communication underlying these platforms amplifies manipulation potential, given the massive scale of processed data, precision, and speed of adaptation (Christiano, 2022). The result is the control of information flow, conditioning of individual autonomy, and erosion of full civic participation.

3.2. Erosion of Public Debate and Democratic Plurality

When analysing the challenges facing public debate and democratic plurality, it is crucial to consider how epistemic vulnerabilities in certain contexts, such as disadvantaged communities, areas with limited educational access, or environments saturated with (dis)information, can impair individuals' capacity for critical thinking and advocacy of their interests (Christiano, 2022). When some citizens lack essential cognitive tools, like access to quality information, critical thinking skills, or reliable communication platforms, while others benefit from contexts that encourage learning, engagement, and access to diverse perspectives, inequality emerges in political sophistication and power dynamics that undermines democracy's foundational principle of equality. This disparity is particularly evident in algorithmic communications, where individuals with more resources dominate discourse, exacerbating pre-existing asymmetries. Consequently, the interplay of conflicting interests and varying political sophistication fosters conditions ripe for manipulation, posing significant threats to vibrant public debate and the consolidation of genuinely pluralist democracy (Christiano, 2022).

However, challenges to public debate and democratic plurality extend beyond information access and political sophistication asymmetries (Couldry & Mejias, 2019). Central to this dynamic is a rising class that, controlling strategic resources such as data and technology, is shaping new power relations.

3.3. Algocracies and the Eclipse of Democratic Judgment

The current concentration of AI dominance lies mostly with the GAFAM (Google, Amazon, Facebook/Meta, Apple, and Microsoft) and China's BAT (Baidu, Alibaba, & Tencent), whose competitive advantage is based on near-exclusive access to massive proprietary data volumes. This setting fosters what some authors term a technological oligopoly characterized not only by limited competition but also by the risk of AI innovation remaining locked in and accessible to society only on terms defined by these companies' commercial interests (Benjamins & Salazar García, 2019). Such concentration of economic and technological power may result in a public sphere subordinated to algorithmic logics geared toward profit maximization at the expense of collective interests (Han 2022). Van Dijck *et al.* (2021) criticise what they call the *platform society*, the role of digital platforms in the organisation of contemporary societies, affecting labour relations, social and civic practices, and affecting democratic processes.

Shoshana Zuboff (2019) describes the emergence of "instrumentarian power" (p. 37), a logic that organizes, aggregates, and standardizes society, replacing politics and democracy with the certainty of computers and social conformity. Zuboff identifies Google and Facebook as pioneers of this model, resisting corporate governance and rejecting democratic demands, securing their claims through economic and political strength. The outcome is a brutally efficient system that reduces space for democratic deliberation, public debate, and individual self-determination, subverting democracy from within and displacing popular sovereignty (Zuboff, 2019).

The term algocracy – government by algorithms – emerges in this context. It denotes a form of governance mediated by algorithms that critically challenges ideals of pluralism, democratic participation, public debate, transparency, and institutional accountability. John Danaher (2016) highlights a problem deeper than mere concealment: opacity. Beyond concealment, which impedes accountability, the opacity of algorithmic systems diminishes society's ability to exercise scrutiny due to the potential incomprehensibility of these processes for human reasoning. Herein lies the problem of legitimacy: growing reliance on algorithmic decisions threatens

public governance not only by limiting accountability and transparency but primarily by making deliberation and scrutiny dependent on processes beyond human understanding (Danaher, 2016).

If indeed we are moving toward a world governed by algorithms, digital literacy ceases to be optional and becomes essential: emerging as a crucial instrument for conscious and effective civic participation. It is in this spirit that the final chapter of this study is approached, where it will be examined how, faced with technological forces potentially greater than ourselves, we can learn to coexist, reflect, and act to ensure that our interactions with AI are grounded in ethics, responsibility, and full participation.

4. Digital Literacy and Citizenship: Ethics, Education, and Balance in the AI Era

Digital literacy is not a mere technical skill; it is a minimal condition for coexistence with AI. In the current context, where algorithms assert themselves beyond mere tools to become agents (Harari, 2025) that shape information flows, perceptions, and political and personal decision-making processes, digital literacy emerges as an essential instrument for intellectual and social survival. It goes beyond mastering digital tools to fostering a proactive digital culture: the capacity to discern facts from misinformation, assess source credibility, and reflect critically on the digital's impact on the public sphere (Fjeld *et al.*, 2020).

Building on this foundation, a reflective exercise follows that considers digital literacy as a critical competence; the role of education in constructing an ethical future; and the balance between security and freedom. These elements will determine how we prepare for a future that is already part of the present.

4.1. Digital Literacy as a Critical Competence in the 21st Century

Accepting that the system described by Zuboff (2019) constitutes “a threat to human nature in the 21st century akin to industrial capitalism’s threat to the natural world in the 19th and 20th centuries” (p. 15), we have an individual and collective moral and ethical obligation to acquire the know-how to navigate this ecosystem safely. Autonomy of choice is not a future projection: today, it inherently depends on our degree of digital literacy.

As the human and digital increasingly merge to become almost inseparable, it is not an exaggeration to state that AI constitutes “a social revolution occurring in real time” (Beam, 2025). From candidate selection to credit eligibility, from the content we consume to judicial decisions, AI’s incorporation into modern life architecture transcends the individual or sectoral sphere. The intellectual fabric of human societies is undergoing profound change as our modes of interaction and thought begin to reflect mechanized logics – with deep effects on empathy (Matzner, 2022), intersubjectivity, and the social bonds that structure community life (Beam, 2025).

This passage captures the essential understanding of digital literacy as a critical, ethical, and educational imperative in the AI era, emphasizing its role in enabling responsible participation and safeguarding autonomy.

It is also important to note that the digital environment operates as a network: unlike traditional television, which flowed only from producer to viewer, digital flows are bidirectional. Every interaction leaves traces; each click, like, or share communicates personal information to content producers; often without the user’s awareness. Understanding these behind-the-screen dynamics is crucial, and users must be fully aware of what they are giving up, notably their privacy, when they are online.

Our model of interaction with AI must evolve from immediate and indiscriminate access to critical, informed, and responsible participation. Digital literacy should be cultivated daily: knowing how to identify

risks, investigate recommendations, verify information accuracy, diversify the AI tools accessed, and reflect on digital habits, beginning to take an active role before technology. Such practices can have a profound impact by transforming citizens into conscious and reflective participants, abandoning the role of mere receivers, consumers, or products. In light of these challenges, the focus shifts to reflecting on how to conceptualize the human-machine relationship.

4.2. Teaching, Ethics, and Democratic Sustainability

The concept of literacy must expand. Learning to read and write is no longer sufficient: integrating the digital component into school curricula from basic to higher education is imperative, establishing it as a permanent teaching element. As Dignum (2023) observes, digital literacy should not remain confined to engineering fields; it must become a transversal element in education, capable of cultivating critical citizens aware of technological impacts and capable of responsible and ethical participation in the digital society.

A lack of proper guidance opens space for distorted learning, exposing individuals to significant risks due to their susceptibility within an ecosystem overloaded with sometimes antagonistic, inaccurate, and fallacious information. Although young people are familiar with technology, they need to develop competencies to critically evaluate available information, identify risks, and act responsibly in the digital environment. Furthermore, invoking the principle of intergenerational solidarity, digital literacy must be accessible to all: unlike digital natives, a significant portion of the global population was not born immersed in technology and requires greater attention to the risks of interconnectedness.

For AI not to become an obstacle to critical thinking, education must go beyond technical transmission and promote literacy that preserves human intellectual autonomy. Initiatives such as AI4People – An Ethical Framework for a Good AI Society (Floridi *et al.*, 2018), the Elements of AI program launched in Finland to democratize technological knowledge (Dignum, 2023), and the work by the Center for Humane Technology reflect efforts toward empowerment that transcend technique, fostering critical reflection and discernment essential for conscious citizenship in the AI era.

What is required, therefore, is not the prohibition of AI use – especially in academic contexts, where restriction may be counterproductive – but rather its integration into education and society, recognizing its inevitability. Failure to do so subjects users to unsupervised learning processes that tend to foster the illusion of choice and the erosion of subjectivity, with the added risk of a significant decline in the intellectual and creative capacities of new generations (Stezhko & Khmil, 2023). As Dignum (2023) notes, this challenge demands diversification of knowledge, incorporating contributions from fields such as philosophy, law, and economics to train citizens and professionals capable of articulating technical skill, ethics, and social responsibility. Unwin (2021) argues that AI should be used ethically in development contexts, emphasising that the focus should not only be on technological innovation, but also on equity, human rights and the empowerment of vulnerable populations. AI can contribute to inclusive and sustainable development.

The key lies in enhancing human intelligence rather than replacing it, following an Intelligence Augmentation logic (Hassani, 2020) whereby AI acts as a partner to our cognitive faculties: technology not as a suppressor of human thought autonomy, but as an amplifier that opens new possibilities for human cognitive performance. From this premise, we advance to reflecting on one of the structuring dilemmas in the AI era, the balance between security and liberty, which constitutes the final topic addressed.

4.3. Balance Between Security and Freedom

This balance is not easily achieved; rather, it is a double-edged sword. The classic tension between security and freedom intensifies as societies digitalize and deepen their integration in an increasingly interconnected world. Technology use for monitoring and surveillance can enhance collective security – through crime prevention, intelligent management of urban systems, or rapid response to emergencies – but simultaneously threatens fundamental freedoms, particularly privacy and individual autonomy, subjecting citizens to forms of permanent control and risks of algorithmic discrimination, whose effects, as noted earlier, produce asymmetrical impacts across social groups.

Digital monitoring and data analysis technologies undoubtedly offer tangible and immediate public safety benefits. Algorithmic surveillance tools contribute to crime prevention, rapid threat detection, and efficient crisis management (Rashid & Kausik, 2024). This promise of greater efficiency and predictability partially explains societal acceptance, especially when associated with protecting essential goods such as life, public safety, or the environment. Nonetheless, this promise carries a structural danger: the normalization of permanent surveillance and the redrawing of boundaries between public and private spheres (Zuboff, 2019). The dilemma is that in the name of security, a control architecture is legitimized that undermines the democratic principles it ostensibly protects.

Today, Nick Bostrom's caution that humanity behaves like children playing with a bomb (2016) resonates with renewed urgency. This vulnerability is reinforced by Harari (2024, 2025), who reminds us that the greatest threat lies not only in AI's potential to create agents more intelligent and creative than ourselves but also in this process reflecting our own nature – competitive, impatient, and often inclined to prioritize progress over safety. Floridi (2019) also reminds us that this trend is not unprecedented: both the climate crisis and the interference of digital platforms in democracy have already exposed risks associated with innovation lacking prior planning.

Benjamins and Salazar García (2019) push this reflection to its hypothetical limit: if there is a deadline to confront the major challenges posed by AI, it may coincide with the so-called Singularity, the moment when machines surpass human intelligence. Even if such a horizon never materializes, it is imperative that human societies develop mechanisms for collective deliberation, with clear concepts and solid frameworks, lest critical decisions affecting everyone are made too late or left to actors indifferent to the common good. For if human power relies on our capacity for large-scale cooperation, it is also this structure that predisposes us to misuse it recklessly (Harari, 2024).

Final Reflections

We live in a historic moment in which the speed and scale of technological innovation, particularly artificial intelligence, threaten to profoundly transform institutions, ways of life, and even the very concept of humanity itself. AI, as a transversal and disruptive technology, is not merely a neutral instrument placed at humanity's disposal; rather, it functions as a mirror of the social, political, and ethical values embedded in those who design, deploy, and govern it. Throughout this work, it has become evident that its potential is deeply ambivalent. On the one hand, AI holds unprecedented capacity to augment human abilities, optimize decision-making, and democratize access to knowledge and services. On the other, it risks amplifying structural inequalities, eroding democratic deliberation, concentrating power in the hands of a few actors, and weakening intellectual autonomy through automation and algorithmic dependence.

The central question, therefore, is not whether humanity should coexist with AI, in fact this coexistence is already an irreversible reality, but rather how this integration will be shaped, regulated, and ethically oriented. The future impact of AI will not be determined by technological sophistication alone, but by the social choices, institutional frameworks, and moral commitments that accompany its development and use.

This analysis has identified three essential and interdependent pillars to guide AI's integration into society.

First, ethics must be understood not as a decorative or secondary consideration, but as the foundational condition under which technological innovation can translate into genuine human progress. Ethical principles such as fairness, accountability, transparency, and respect for human dignity must be embedded into AI systems from their conception, rather than retrofitted after harm has occurred (Han, 2015). Without this ethical grounding, innovation risks becoming detached from human values and social responsibility.

Second, regulation emerges not as an obstacle to technological advancement, but as an indispensable common framework capable of aligning innovation with democratic ideals. Effective regulation can prevent abuses, reduce asymmetries of power, and ensure that AI serves collective interests rather than narrow economic or political agendas. In this sense, regulation should be flexible enough to encourage innovation while robust enough to protect fundamental rights, striking a delicate balance between security and freedom, progress and accountability.

Finally, digital literacy must be recognized as a core condition of modern citizenship. It goes far beyond technical proficiency, encompassing critical understanding of how algorithms function, how data is used, and how technological systems influence perception, behaviour, and decision-making. Without this literacy, individuals are more susceptible to manipulation, misinformation, and passive consumption of algorithmically curated realities, undermining their capacity for autonomous judgment and meaningful participation in democratic life.

Ultimately, AI will neither redeem nor doom humanity; it will simply reflect the values, priorities, and power structures of the societies that create and govern it. If driven primarily by competition, surveillance, and short-term profit, it will reproduce and intensify these logics. If, however, it is guided by cooperation, transparency, inclusivity, and responsibility, AI can become a powerful catalyst for a more just, sustainable, and humane future. Preserving and expanding our concept of humanity in the age of artificial intelligence requires the deliberate cultivation of collective intelligence, one that harmonizes technological innovation with human subjectivity, creativity, and ethical judgment. In the end, it is not algorithms that will determine humanity's fate, but the consciousness and values that guide our choices. Ethics, therefore, must lead the way not as a constraint on technology, but as its most essential compass.

Bibliography

- American Psychological Association. (2018). Intelligence, in *APA Dictionary of Psychology*. <https://dictionary.apa.org/intelligence>
- Baerren, E. (2025). *What happens if artificial intelligence becomes self-aware*. Central Michigan University. <https://www.cmich.edu/news/details/what-happens-if-artificial-intelligence-becomes-self-aware>
- Beam, B. (2025, April 2). *The social price of AI communication*. IE University. <https://www.ie.edu/insights/articles/the-social-price-of-ai-communication/>

- Bellman, R. (1957). A Markovian decision process. *Journal of Mathematics and Mechanics*, 6(5), 679-684. <https://doi.org/10.1512/iumj.1957.6.56038>
- Bellman, R. (1978). *An introduction to artificial intelligence: Can computers think?* Boyd & Fraser Publishing Company.
- Benjamins, V. R., & Salazar García, I. (2019). Towards a framework for understanding societal and ethical implications of artificial intelligence, in I. B. Alfonso, L. M. Fernández-Martínez, & R. Suárez-Álvarez (eds.), *Vulnerabilidad y cultura digital: Riesgos y oportunidades de la sociedad hiperconectada* (pp. 89-100). Dykinson. <https://doi.org/10.48550/arXiv.2001.09750>
- Bostrom, N. (2016, June 12). Artificial intelligence: "We're like children playing with a bomb." *The Guardian*. <https://www.theguardian.com/technology/2016/jun/12/nick-bostrom-artificial-intelligence-machine>
- Christiano, T. (2022). Algorithms, manipulation, and democracy. *Canadian Journal of Philosophy*, 52(1), 109-124. <https://doi.org/10.1017/can.2021.29>
- Couldry, N., & Mejías, U. A. (2019). *The costs of connection: How data is colonizing human life and appropriating it for capitalism*. Stanford University Press.
- Crume, J. (2024). *AI, machine learning, deep learning and generative AI explained* [Video]. YouTube. <https://www.youtube.com/watch?v=qYNweeDHiyU>
- Damásio, A. (2000). *O sentimento de si: O corpo, a emoção e a neurobiologia da consciência*. Publicações Europa-América.
- Danaher, J. (2016). The threat of algocracy: Reality, resistance and accommodation. *Philosophy & Technology*, 29(3), 245-268. <https://doi.org/10.1007/s13347-015-0211-1>
- Daylight, E. G. (2015). Towards a historical notion of Turing – the father of computer science. *History and Philosophy of Logic*, 36(3), 205-228. <https://doi.org/10.1080/01445340.2015.1082050>
- Dignum, V. (2023). Responsible artificial intelligence: Recommendations and lessons learned, in D. O. Eke, K. Wakanuma, & S. Akintoye (eds.), *Responsible AI in Africa* (pp. 195-214). Palgrave Macmillan. https://doi.org/10.1007/978-3-031-08215-3_9
- European Commission. (2019). *Ethics guidelines for trustworthy AI*. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- European Commission Directorate-General for Communications Networks, Content and Technology. (2020). *The assessment list for trustworthy artificial intelligence (ALTAI) for self-assessment*. Publications Office of the European Union. <https://data.europa.eu/doi/10.2759/002360>
- Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). *Principled artificial intelligence: Mapping consensus in ethical and rights-based AI frameworks*. Berkman Klein Center for Internet & Society, Harvard University.
- Floridi, L. (2019). Establishing the rules for building trustworthy AI. *Nature Machine Intelligence*, 1(6), 261-262. <https://doi.org/10.1038/s42256-019-0055-y>
- Floridi, L. (2022). *Ethics of artificial intelligence: Principles, challenges and opportunities*. Oxford University Press.

- Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., & Vayena, E. (2018). AI4People – An ethical framework for a good AI society. *Minds and Machines*, 28(4), 689-707. <https://doi.org/10.1007/s11023-018-9482-5>
- Han, B.-C. (2015). *The transparency society*. Stanford University Press.
- Han, B.-C. (2022). *Infocracy: Digitalization and the crisis of democracy*. Polity Press.
- Harari, Y. N. (2023, April 3). *Post on X*. https://x.com/harari_yuval/status/1642814487551000576
- Harari, Y. N. (2024, August 24). Never summon a power you can't control: Yuval Noah Harari on how AI could threaten democracy and divide the world. *The Guardian*. <https://www.theguardian.com/technology/article/2024/aug/24/yuval-noah-harari-ai-book-extract-nexus>
- Harari, Y. N. (2025). *AI and human evolution* [Video]. YouTube. <https://www.youtube.com/watch?v=jt3UI3rPXaE>
- Hassani, H., Silva, E. S., Unger, S., TajMazinani, M., & Mac Feely, S. (2020). Artificial intelligence or intelligence augmentation: What is the future? *AI*, 1(2), 143-155. <https://doi.org/10.3390/ai1020008>
- Krafft, P. M., Young, M., Katell, M., Huang, K., & Buggingo, G. (2020). Defining AI in policy versus practice, in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 195-201). Association for Computing Machinery. <https://doi.org/10.1145/3375627.3375835>
- Matzner, T. (2022). Digital infrastructures and the erosion of empathy. *Media, Culture & Society*, 44(7), 1321-1339.
- Pariser, E. (2011). *Beware online "filter bubbles"* [Video]. YouTube. <https://www.youtube.com/watch?v=B8ofWFx525s>
- Rai, D. H. (2024). Artificial intelligence through time: A comprehensive historical review. ResearchGate. <https://doi.org/10.13140/RG.2.2.22835.03364>
- Rashid, A. B., & Kausik, M. A. K. (2024). AI revolutionizing industries worldwide. *Hybrid Advances*, 7, Article 100277. <https://doi.org/10.1016/j.hybadv.2024.100277>
- Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (3rd ed.). Pearson. <https://people.engr.tamu.edu/guni/csce625/slides/AI.pdf>
- Smith, C., McGuire, B., Huang, T., & Yang, G. (2006). *The history of artificial intelligence*. University of Washington. <https://courses.cs.washington.edu/courses/csep590/06au/projects/history-ai.pdf>
- Stanford Encyclopedia of Philosophy. (2014). *Consciousness*. Stanford University. <https://plato.stanford.edu/entries/consciousness/>
- Stezhko, Z. V., & Khmil, T. V. (2023). Artificial intelligence as a socio-cultural phenomenon. *Anthropological Measurements of Philosophical Research*, 24, 68-74. <https://doi.org/10.15802/ampr.v0i24.295317>
- Tallberg, J., Erman, E., Furendal, M., Geith, J., Klamberg, M., & Lundgren, M. (2023). The global governance of artificial intelligence. *International Studies Review*, 25(3). <https://doi.org/10.1093/isr/viad040>
- Unwin, T. (2021). Towards ethical AI for development. *Information Technologies & International Development*, 17(1), 25-39.
- Van Dijck, J., Poell, T., & de Waal, M. (2021). *The platform society: Public values in a connective world*. Oxford University Press.

Vasdani, T. (2020). *Robot justice: China's use of internet courts*. Law360 Canada. <https://www.law360.ca/ca/articles/1750396/robot-justice-china-s-use-of-internet-courts>

Yordanova, K., & Bertels, N. (2024). Regulating AI: Challenges and the way forward, in H. Sousa Antunes *et al.* (eds.), *Multidisciplinary perspectives on artificial intelligence and the law* (pp. 441-456). Springer. https://doi.org/10.1007/978-3-031-41264-6_23

Zuboff, S. (2019). *The age of surveillance capitalism: The fight for a human future at the new frontier of power*. Harvard University Press.

Recebido para publicação: 24 de novembro de 2025

Aceite após revisão: 17 de dezembro de 2025

Received for publication: 24 November 2025

Accepted in revised form: 17 December 2025