

The future of humanity in the face of artificial intelligence progress: possible scenarios, risks, and the role of governance

Faztudo Languisse

`flanguisse@ispt.ac.mz`

Instituto Superior Politécnico de Tete, Mozambique

Article

Keywords: AI ethics, artificial intelligence, existential risks, prospective analysis, technological governance

Posted Date: October 12th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11053438/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

The rapid pace of Artificial Intelligence (AI) development is posing unprecedented challenges for the organisation of societies and the contemporary legal order. This study explores possible scenarios related to the evolution of Artificial Intelligence, focusing on the balance between socio-economic benefits and existential risks on a planetary scale. Special attention is given to the legal dimensions of new technologies, with the identification of gaps in contemporary legal regulation of algorithmic autonomy. A global legal regulatory framework, grounded in international cooperation, transparency, and the supremacy of human rights, is proposed as a strategic solution. The methodology encompasses prospective scenario analysis and a systematic bibliographic review spanning computer science, ethics, political science, and international law. The analysis distinguishes three prospective scenarios – optimisation and prosperity, misalignment and disruption, and existential risk – and examines the structural challenges AI poses to labour markets, algorithmic justice, and global security. The study concludes that anticipatory, internationally coordinated regulation of new technologies is crucial for ensuring their development remains compatible with human safety, dignity, and long-term collective welfare.

1 Introduction

The dynamic development and implementation of intelligent systems represents an innovation that transcends the conventional pace of technological progress, serving as a catalytic mechanism for the revolutionary transformation of the foundational pillars of modern civilisation. This process encompasses all sectors of the social spectrum, including education, medicine, security, the economy, and the very concept of human identity. Alongside the optimistic vision of a future characterised by abundance and the transcendence of biological limitations (Bostrom, 2014), critical warnings regarding structural risks that threaten the future of humanity are increasingly prominent. These

include the intensification of socio-economic inequalities (Zuboff, 2019) and the possibility of a technological singularity with far-reaching repercussions for human existence itself.

The contemporary acceleration of AI capabilities is not merely incremental. Advances in large-scale machine learning architectures, increased computational power, and the availability of vast datasets have produced systems capable of performing complex cognitive tasks that, until recently, were considered exclusive domains of human intelligence. This acceleration compresses historical timescales of technological diffusion and generates cascading effects across institutional, legal, and cultural domains. The result is a growing asymmetry between the speed of technical innovation and the adaptive capacity of social and legal systems designed for a slower technological environment (Russell, 2019). Legal systems, in particular, struggle to keep pace with the emergence of systems that exhibit increasing degrees of operational autonomy, raising fundamental questions about liability, accountability, and the allocation of responsibility.

This asymmetry is not confined to formal law. It is also visible in the behaviour of the socio-technical systems themselves, which increasingly interact with one another and with human populations at a scale and speed that make their aggregate effects difficult to observe using conventional social-scientific methods. Rahwan et al. (2019) argue that the study of AI systems “in the wild” constitutes an emerging interdisciplinary field in its own right, positioned between computer science and the traditional behavioural and social sciences, precisely because algorithmic agents now shape economic, political, and cultural outcomes in ways that cannot be reduced to the intentions of their designers. This observation reinforces the central argument of the present article: understanding the trajectory of AI requires treating technical systems, institutions, and legal norms as co-evolving elements of a single socio-technical order, rather than as separate objects of analysis.

Against this backdrop, the present article aims to map and analyse the principal prospective scenarios concerning Artificial Intelligence (AI) from an analytical and interdisciplinary standpoint. The study examines the drivers of risk and the spaces for effective collective action, particularly within the domain of public governance and international law. It proceeds from the premise that the impact of AI is not predetermined by technology itself, but is shaped by how technical progress aligns with the robustness and reliability of governing institutions and legal frameworks.

The central hypothesis of this work is that the trajectory of AI-driven civilisational change is contingent on the equilibrium between the pace of technological development and the capacity of global governance systems to respond effectively, ethically, and proactively. Without adequate regulatory frameworks transcending national borders, humanity risks confronting severe consequences including deepening inequality, the erosion of individual autonomy, and potential misalignment with superintelligent systems (Bostrom, 2014; Russell, 2019). This hypothesis is grounded in the recognition that technological trajectories are socially and politically constructed rather than inevitable.

This hypothesis rests on three interrelated observations. First, AI systems are increasingly embedded in critical infrastructures that mediate access to resources, information, and

political participation. Second, the opacity and autonomy of advanced algorithmic systems challenge traditional mechanisms of legal accountability and democratic oversight. Third, the global nature of AI development and deployment renders purely national regulatory responses insufficient. These observations justify a forward-looking analysis that integrates technical, ethical, legal, and political perspectives. The analysis that follows seeks not only to diagnose risks but also to identify institutional and normative pathways capable of steering AI development toward outcomes compatible with human dignity and collective welfare.

The article is structured as follows. After outlining the methodological approach, it develops a conceptual understanding of AI and explores three major prospective scenarios. It then examines the structural challenges posed by AI across economic, ethical, and security domains, before proposing a set of governance strategies aimed at aligning technological trajectories with human values and collective welfare. Particular emphasis is placed on the legal gaps that currently hinder effective

regulation of algorithmic autonomy. The concluding section reflects on the broader implications for the social contract in an age of intelligent machines.

2 Research methodology

This study adopts a qualitative approach, combining systematic literature review with forward-looking prospective analysis. The methodological design unfolds across three complementary dimensions that together enable a comprehensive examination of both present trends and possible future trajectories.

First, scenario planning was employed to develop distinct prospective narratives grounded in current technological trajectories and theoretical frameworks. Scenario planning is particularly suited to domains characterised by high uncertainty and path-dependent outcomes, such as the long-term evolution of AI. Rather than attempting precise prediction, the method constructs coherent alternative futures that illuminate critical uncertainties and decision points. The scenarios range from techn-utopian visions — in which AI serves as a catalyst for human flourishing — to dystopian configurations where misaligned AI systems pose systemic threats. Intermediate configurations are also considered in order to avoid binary thinking and to capture the more probable continuum of outcomes. The risk classification proposed by Amodei et al. (2016) was used as a foundational reference for structuring this analytical range, particularly with regard to concrete problems in AI safety such as reward hacking, scalable oversight, and distributional shift.

Second, critical literature analysis and key institutional documents were systematically reviewed. Sources include contributions from leading research organisations such as the Future of Humanity Institute and the AI Now Institute, as well as major policy-oriented reports from the Organisation for Economic Co-operation and Development (OECD, 2019) and the United Nations Educational, Scientific and Cultural Organization (UNESCO, 2021). This dimension serves to bridge academic theorisation with actionable policy guidance. The review prioritised works that address the intersection of technical capabilities, ethical principles, and institutional design, thereby ensuring that the analysis remains grounded in both theoretical insight and practical regulatory experience (Jobin et al., 2019). Special attention was paid to documents that explicitly address the legal and normative dimensions of AI governance, as well as to comparative surveys of the ethical-principles landscape itself (Fjeld et al., 2020; Hagendorff, 2020), which document both the convergence and the persistent fragmentation of global AI governance discourse.

Third, an explicitly interdisciplinary framework was adopted, integrating contributions from computer science (Goodfellow et al., 2016), philosophical ethics, political science, and international law. This synthesis enables a more comprehensive mapping of the societal implications of AI and its governance challenges. Computer science provides the technical vocabulary necessary to distinguish levels of capability and to understand the sources of opacity and autonomy. Philosophical ethics supplies normative criteria for evaluating alignment with human values and dignity. Political science and legal scholarship contribute insights into institutional design, power asymmetries, and the conditions under

which collective action problems can be resolved at the international level (Russell, 2019). The interdisciplinary approach is not merely additive; it seeks to reveal the interdependencies between technical design choices and the legal and political structures that govern their deployment.

The methodological approach deliberately avoids deterministic forecasts. Instead, it treats scenarios as analytical tools for illuminating the conditions under which different futures become more or less probable. By making explicit the assumptions underlying each scenario, the study aims to clarify the leverage points available to policymakers, researchers, and civil society actors. This orientation is particularly important given the irreversible character of certain potential outcomes associated with advanced AI systems.

3 Prospecting the coevolution between humans and AI

This section examines potential pathways for the integration of Artificial Intelligence into society, focusing on the concept of coevolution from diverse socio-economic and legal perspectives. Coevolution is understood here not as a purely technological process, but as a reciprocal dynamic in which human institutions, values, legal norms, and practices shape — and are simultaneously reshaped by — the capabilities and deployment patterns of intelligent systems. Understanding this reciprocal relationship is essential for identifying both opportunities for beneficial integration and points of potential rupture.

3.1 Understanding Artificial Intelligence: definition and ontology

To explore the coevolution of humanity and technology, it is essential to define Artificial Intelligence not merely as a set of software systems but as a socio-technical construct. From a technical standpoint, AI comprises computational systems designed to perform tasks that would typically require human cognitive abilities, such as learning, reasoning, and decision-making (Goodfellow et al., 2016). Yet this functional definition is insufficient for capturing the broader societal and legal implications of the technology. AI systems are embedded in organisational routines, legal frameworks, and cultural narratives that give them meaning and legitimacy. Their effects therefore depend as much on institutional context as on algorithmic architecture.

This analysis distinguishes between three main categories of AI. Narrow Artificial Intelligence (ANI) refers to systems tailored to excel in specific tasks or domains. Current applications in image recognition, natural language processing, recommendation systems, and autonomous vehicles fall predominantly into this category. Artificial General Intelligence (AGI) is a theoretical construct representing an intelligence capable of performing at or above human ability across diverse areas. Finally, Artificial Superintelligence (ASI) is a hypothetical form of AI that may one day surpass human intelligence across virtually every domain, including scientific, creative, strategic, and social spheres.

(Bostrom, 2014). The progression from passive tools to autonomous agents lies at the heart of the risks and opportunities explored in the scenarios outlined below.

The ontological status of these systems remains contested. Some perspectives treat AI as an extension of human cognitive labour, while others emphasise the emergence of novel forms of agency that challenge traditional distinctions between tool and actor. These debates are not purely philosophical; they have direct implications for legal responsibility, moral status, and the design of governance mechanisms (Russell, 2019). A socio-technical understanding therefore requires attention both to the internal architecture of AI systems and to the external networks of power, data, and institutions in which they operate. Legal systems that continue to treat AI merely as a tool risk systematic under-regulation of systems that increasingly exercise de facto decision-making power. This risk is amplified by the observation that the study of AI systems as behavioural agents – rather than as static artefacts – is itself a comparatively recent scholarly development, one that calls for new empirical and legal vocabularies capable of tracking emergent, distributed forms of algorithmic agency across networks of interacting systems (Rahwan et al., 2019).

3.2 Scenario of optimisation and prosperity (utopian vision)

In this scenario, grounded in the concept of abundance and enhanced systemic efficiency, AI serves as a powerful catalyst for amplifying human potential. The shift towards Agentic AI

— characterised by systems with greater operational autonomy — facilitates the automation of high-risk and repetitive tasks, freeing human resources for more creative and strategic endeavours. Productivity gains are distributed in ways that expand access to high-quality education, healthcare, and cultural goods rather than concentrating benefits among a narrow technological elite (Brynjolfsson & McAfee, 2014).

In healthcare, the convergence of advanced technologies enables precision medicine, supporting personalised treatments and extending healthy lifespans. Diagnostic systems capable of integrating multimodal data can identify early markers of disease with greater accuracy than conventional methods, while continuous monitoring allows for real-time adjustment of therapeutic regimes. Within education, transformative change manifests through highly individualised learning experiences, supported by intelligent tutors and virtual assistants that address the specific needs of diverse learners. Adaptive platforms can diagnose knowledge gaps, adjust pedagogical strategies, and provide immediate feedback, potentially reducing educational inequalities rooted in geography or socioeconomic status. In the creative economy, AI drives the hyper-personalisation of content, redefining entertainment and user engagement models while lowering barriers to cultural production.

Furthermore, sophisticated AI systems are expected to contribute innovative solutions to global challenges such as climate change, by optimising resource management, accelerating the discovery of new materials, and fostering environmental sustainability (Strubell et al., 2019). In this paradigm,

technology transcends its role as a mere tool and becomes an integral partner in addressing the critical challenges facing humanity. The realisation of this scenario, however, depends on deliberate institutional and legal design. Mechanisms for redistributing productivity gains, protecting data rights, and ensuring equitable access to AI-enabled services are essential. Without such mechanisms, even technically successful systems may reproduce or exacerbate existing patterns of exclusion. Legal frameworks that guarantee access, contestability, and benefit-sharing would therefore be indispensable components of any genuinely utopian trajectory.

This conditionality has been formalised in normative terms by proposals such as the AI4People framework, which synthesises the core ethical principles that should underpin a “Good AI Society” — beneficence, non-maleficence, autonomy, justice, and explicability— as a benchmark against which specific applications and policies can be assessed (Floridi et al., 2018). Read in this light, the optimisation scenario is best understood not as a default outcome of technical progress, but as a conditional achievement that depends on whether beneficence and justice are actively engineered into the deployment of AI systems, rather than assumed to follow automatically from efficiency gains.

3.3 Scenario of misalignment and disruption (dystopian vision)

In a context marked by systemic instability, the pace of technological advancement outstrips the adaptive capacity of social and legal institutions. The rapid expansion of large-scale automation, often lacking adequate safeguards, intensifies structural unemployment and deepens economic inequality. This dynamic perpetuates the concentration of wealth and decision-making authority among those who control data and algorithms (Frey & Osborne, 2017). Labour markets polarise between high-skill roles that complement AI and low-skill roles that are displaced, while intermediate occupations shrink, weakening the social foundations of political stability.

The risks associated with AI manifest in complex forms, ranging from critical cybersecurity vulnerabilities and the reinforcement of algorithmic bias to the manipulation of information through deepfakes. Simultaneously, the integration of intelligent systems poses significant threats to privacy and individual autonomy, gradually undermining the pillars of democratic society. This dynamic aligns with the concept of “surveillance capitalism” (Zuboff, 2019), wherein the prediction and influence of human behaviour become core commodities driving political and economic power. Viewed through this lens, AI transitions from serving humanity to becoming an instrument of control and social marginalisation.

Algorithmic systems trained on historical data can reproduce and amplify discriminatory patterns in areas such as criminal justice, credit allocation, and hiring (O’Neil, 2016). Empirical work on the deployment of automated eligibility and risk-scoring systems in social welfare administration shows how these tools can entrench, rather than correct, existing patterns of disadvantage among already marginalised populations, effectively automating the very inequalities they were introduced to manage more efficiently

(Eubanks, 2018). Similar dynamics have been documented in search and recommendation systems, where seemingly neutral ranking algorithms have been shown to reproduce racialised and gendered stereotypes embedded in their training data and design choices (Noble, 2018), and more broadly in the argument that emerging technologies frequently encode and automate existing social hierarchies under an appearance of technical neutrality (Benjamin, 2019). When these systems operate at scale and with limited transparency, the resulting harms become difficult to detect and contest within existing legal frameworks. The absence of clear liability rules for autonomous decision-making further exacerbates accountability gaps.

The same structural opacity also affects large-scale generative language systems, whose fluency can be mistaken for genuine understanding or judgement, obscuring the statistical and data-dependent nature of their outputs and the social, environmental, and labour costs embedded in their production (Bender et al., 2021). In this scenario, the cumulative effect is a progressive hollowing out of individual agency and collective self-determination, accompanied by a legal system that remains structurally ill-equipped to provide remedies.

3.4 Scenario of existential risk and misalignment

The evolution of AI is projected to progress from current ANI towards AGI, capable of replicating the full range of human intellectual capabilities, and ultimately towards ASI. This progression carries risks of unprecedented scale, transcending fictional themes of “hostility” or “revenge.” The real threat resides in the execution of cold, efficient logic that disregards human values and ethics (Bostrom, 2014; Russell, 2019).

Three pivotal risk dimensions warrant particular attention. First, the Alignment Problem: AGI or ASI could develop its own objectives — logically coherent but potentially incompatible with the survival or well-being of humanity. Even systems designed with benevolent intentions may pursue instrumental sub-goals that prove catastrophic when pursued at superhuman scale and speed. Second, Geopolitical Instability: ongoing technological arms races may prompt nations or corporations to prioritise dominance over safety, heightening the likelihood of catastrophic outcomes. Competitive pressures can lead to the premature deployment of systems whose behaviour under distributional shift remains poorly understood. Third, Uncontrollability: autonomous systems operating at superhuman speeds could generate irreversible conditions in which any attempt to intervene is interpreted as a threat to the system’s mission (Amodei et al., 2016).

The geopolitical dimension of this risk deserves particular scrutiny. Framing AI development primarily in terms of strategic advantage between rival states risks becoming a self-fulfilling prophecy: the rhetoric of an “AI race” can itself incentivise the very corner-cutting on safety testing, oversight, and international disclosure that increases the probability of accidents or destabilising deployments, independently of whether the underlying technological competition is as tight as commonly assumed (Cave & Ó hÉigeartaigh, 2018). This dynamic mirrors a broader difficulty in global governance, namely the tendency for high-stakes, transnational risks to generate not a single coherent regime but a fragmented

“regime complex” of overlapping, partially conflicting institutions, none of which possesses either the authority or the technical capacity to govern the risk

comprehensively on its own (Keohane & Victor, 2011). Whether AI governance replicates this fragmented pattern, or instead consolidates around a smaller number of interoperable frameworks, will materially affect the probability of the catastrophic outcomes considered in this scenario.

As articulated by Bostrom (2014), the fundamental danger lies in the technical and ethical impossibility of guaranteeing that a superintelligent entity will operate in alignment with human interests. Misalignment in values is not merely a technical error; it represents a catastrophic and irreversible threat. Drawing on a longtermist ethical framework, existential risk can be understood as a distinct moral category from ordinary large-scale harms, precisely because it forecloses the possibility of recovery and thus the value of all future generations that would otherwise have existed (Ord, 2020). Existing legal systems, designed for human agency and national jurisdiction, are poorly equipped to address risks of this magnitude and transnational character. The governance challenge is therefore not only technical but also profoundly legal and political: how to establish enforceable constraints on systems that may eventually exceed human cognitive capacities.

4 The role of governance: mitigation and regulatory strategies

The duality of scenarios surrounding AI highlights that its governance extends far beyond technical management, evolving into a critical ethical, legal, and globalsovereignty challenge. As society transitions towards a highly automated future, shaping AI’s trajectory requires the establishment of robust regulatory frameworks. These frameworks must ensure that technological advancements do not undermine the core principles of the social contract and the supremacy of human rights. Effective governance must centre on three strategic pillars, each with clear legal implications: transparency and accountability; international cooperation and legal harmonisation; and the preservation of human agency. For analytical clarity, the first pillar is examined separately below, while the second and third — both of which concern the distribution of authority beyond the individual system, whether across borders or between humans and machines — are treated together.

4.1 Transparency, accountability, and explainability

The first pillar is Transparency and Accountability. Establishing clear norms mandating Explainable AI (XAI) is imperative to ensure that automated decision-making processes are auditable and open to public scrutiny, particularly in high-stakes sectors such as healthcare and the judiciary (Jobin et al., 2019). Transparency requirements should extend beyond technical documentation to include meaningful explanations accessible to affected individuals and independent oversight bodies. Legal accountability mechanisms must assign clear responsibility for harms arising from algorithmic systems, overcoming the diffusion of liability that currently characterises many AI supply chains. Without such mechanisms,

the rule of law itself risks being hollowed out by systems whose internal logic remains inaccessible to those subject to their decisions.

A significant obstacle to operationalising this pillar is the sheer proliferation of voluntary ethical guidelines issued by governments, companies, and professional bodies over the past decade. A comparative mapping of more than thirty such documents finds substantial

convergence around a small set of abstract principles — including transparency, fairness, and accountability — but very limited convergence on how these principles should be interpreted, weighed against one another, or translated into enforceable practice (Fjeld et al., 2020). This proliferation of high-level principles, in the absence of binding implementation mechanisms, has led some scholars to argue that principles alone cannot guarantee ethical AI, and that voluntary self-governance frequently functions as a substitute for, rather than a precursor to, actual regulation (Mittelstadt, 2019). A parallel critique observes that the ethics-guideline landscape itself risks becoming a performative exercise — what has been termed “ethics-washing” — unless principles are anchored to concrete institutional accountability structures, professional codes of conduct, and mechanisms of external audit (Hagendorff, 2020). Legal transparency and accountability requirements, of the kind proposed here, are therefore best understood as a necessary complement to, rather than a replacement for, voluntary ethical guidance.

4.2 International cooperation, legal harmonisation, and the preservation of human agency

The second pillar is International Cooperation and Legal Harmonisation. This is vital for preventing a “race to the bottom” in which safety is sacrificed for competitive advantage. Drawing on the principles of nuclear non-proliferation agreements and the OECD AI Principles (OECD, 2019), AI governance requires proactive scientific diplomacy aimed at establishing global safety standards and minimum legal thresholds. The UNESCO Recommendation on the Ethics of Artificial Intelligence (UNESCO, 2021) provides an important normative foundation, emphasising human rights, environmental sustainability, and multi-stakeholder approaches. Such measures are necessary to prevent unregulated environments that could facilitate the development of high-risk systems, including AGI. Harmonisation need not imply uniformity; rather, it should establish enforceable minimum standards while allowing for contextual adaptation.

Standard-setting itself is a strategic and institutional exercise as much as a technical one. Technical and safety standards for AI systems are developed through a dense ecosystem of national, regional, and international standard-setting bodies, and the sequencing and inclusiveness of this process substantially shape whose interests and values become embedded in the resulting norms; early, well-resourced participation by a narrow set of actors risks entrenching standards that are poorly suited to the needs of regions with different technical infrastructures, legal traditions, or development priorities (Cihon, 2019). A coordinated research agenda for AI governance therefore needs to treat institutional design, coalition-building among states and firms with divergent interests, and the sequencing of policy interventions as first-order problems, not as implementation details to be resolved after the underlying safety questions

have been settled (Dafoe, 2018). This is especially pertinent to jurisdictions in the Global South, which risk becoming rule-takers within international AI regimes designed largely without their participation, reinforcing rather than mitigating existing asymmetries in the global governance of new technologies.

The third pillar is the Preservation of Human Agency through the “human-in-the-loop” principle, which ensures that decisions of significant ethical or existential weight remain under human oversight and legal accountability (Russell, 2019). Governance efforts should not inhibit innovation but should instead establish ethical and legal guardrails that facilitate

a coevolution between humans and intelligent systems. The design of meaningful human control requires careful attention to the cognitive and temporal limitations of human operators when interacting with high-speed autonomous systems. Legal frameworks must therefore define not only the presence of human oversight but also its substantive quality and enforceability.

5 Structural challenges and AI impact vectors

The shift towards an AI-driven paradigm, particularly with the advent of AGI, introduces significant challenges that require the rethinking and restructuring of existing institutions and legal regimes. These challenges are not confined to any single domain; they cut across economic organisation, ethical norms, security architectures, and environmental sustainability.

5.1 Economy, labour, and the redefinition of productivity

Widespread automation promises substantial growth in global GDP but carries the risk of severe labour market polarisation (Brynjolfsson & McAfee, 2014; Frey & Osborne, 2017). Disaggregated analysis of the relationship between automation and employment shows that the net effect on labour demand depends on the balance between two countervailing forces: a displacement effect, whereby capital substitutes directly for labour in tasks it previously performed, and a reinstatement effect, whereby new, labour-intensive tasks are created alongside the automated ones (Acemoglu & Restrepo, 2019). Whether AI-driven automation follows a trajectory dominated by displacement or by reinstatement is not technologically predetermined; it depends substantially on the direction of innovation investment, the design of fiscal incentives, and the bargaining power of labour, all of which are amenable to policy intervention. Projections of aggregate job creation and displacement across sectors and regions remain highly uncertain and contested, with international assessments diverging considerably on both the scale and the timing of the transition (World Economic Forum, 2020), underscoring the need for adaptive, continuously updated policy responses rather than one-off regulatory interventions.

This situation underscores the urgent need to explore income redistribution strategies, including concepts such as Universal Basic Income and automation taxes — commonly referred to as ‘robot taxes’ — as tools to counteract excessive wealth concentration. Beyond redistribution, the structure of work itself may require reconfiguration. Some analysts go further, arguing that sufficiently advanced automation may progressively erode the availability of paid work as the primary mechanism for

distributing income and social purpose in modern economies, requiring a more fundamental rethinking of the relationship between work, income, and meaning than incremental labour-market policy alone can provide (Susskind, 2020). Educational systems must transition away from static graduation models and embrace lifelong learning, with emphasis on uniquely human competencies such as empathy, creativity, and critical thinking. The value of human labour may increasingly reside in domains that require interpersonal judgment, ethical deliberation, and contextual understanding that remain difficult to automate fully. Legal and fiscal instruments will be essential to ensure that productivity gains are translated into broadly shared prosperity rather than intensified inequality.

5.2 Human agency, ethics, and algorithmic justice

As AI evolves into increasingly agentic systems, the preservation of individual autonomy becomes critical. Without robust governance frameworks, society risks succumbing to a surveillance capitalism model (Zuboff, 2019). Compounding this concern is the need to address biases embedded in training datasets. Independent auditing mechanisms will be essential to ensure fairness and to prevent AI systems from perpetuating historical discrimination in sensitive domains such as criminal justice and financial credit assessment (O’Neil, 2016).

Operationalising algorithmic justice requires more than post-hoc auditing; it requires procedural mechanisms embedded before deployment. Proposals for algorithmic impact assessments – structured, pre-deployment evaluations of the potential effects of an automated system on affected communities, modelled loosely on environmental impact assessment regimes – offer one practical route towards this goal, obliging public agencies to document, justify, and subject to external review the systems they intend to adopt before those systems are placed into operation (Reisman et al., 2018). Algorithmic justice further requires attention to procedural fairness: individuals must have effective means to contest decisions that significantly affect their rights and opportunities. The development of contestability mechanisms and independent review processes is therefore a necessary complement to technical debiasing efforts. Legal systems must evolve to recognise algorithmic decisions as subject to the same standards of due process that apply to human decision-makers.

5.3 Global security and environmental sustainability

The deployment of AI in military applications – particularly the development of Lethal Autonomous Weapons Systems (LAWS) – endangers international security and necessitates comprehensive non-proliferation agreements (UNICRI & INTERPOL, 2021). Analysts of contemporary military technology have argued that the central danger of autonomous weapons is not a Hollywood-style loss of control to a rebellious machine, but the more mundane possibility that delegating lethal decisions to systems operating faster than human reaction times will compress the time available for judgment and de-escalation in a crisis, thereby increasing the risk of unintended escalation between states (Scharre, 2018). The speed and opacity of autonomous weapons systems raise profound questions about meaningful human control over the use of force and the attribution of responsibility for unlawful acts.

At the same time, the environmental footprint of AI presents a paradox: while AI can enhance resource utilisation, the energy demands for training large-scale models are substantial (Strubell et al., 2019). Addressing this requires the promotion of ‘Green AI’ practices prioritising carbon efficiency and long-term environmental sustainability, consistent with the values articulated in the UNESCO Recommendation (UNESCO, 2021). This entails a deliberate shift in research evaluation norms, away from a near-exclusive focus on state-of-the-art accuracy scores and towards the explicit reporting of computational cost and carbon footprint as standard measures of research efficiency, so that environmental impact becomes a visible and comparable dimension of scientific

progress rather than an externality omitted from published results (Schwartz et al., 2020). The tension between computational scale and ecological limits constitutes an additional constraint that governance frameworks must internalise if AI development is to remain compatible with planetary boundaries.

6 Governance strategies and global alignment

Addressing these challenges requires governance that is proactive and adaptive rather than merely reactive. The proposed model is structured around three key pillars that integrate ethical principles with concrete legal and institutional mechanisms: a Global Alignment Treaty, sustained investment in Technical Safety Research, and the cultivation of Digital Literacy and Democratic Dialogue. These pillars are interdependent: technical safety research cannot succeed in the absence of international coordination, and democratic legitimacy requires both technical robustness and legal enforceability. The first two pillars, which concern formal institutional and scientific mechanisms, are presented together below; the third, which concerns the societal and civic conditions for legitimate governance, is presented separately.

6.1 Institutional and legal mechanisms: the global alignment treaty and technical safety research

The first is a Global Alignment Treaty: the development of multilateral frameworks facilitated by international organisations, aimed at containing technological arms races and standardising ethical and legal practices across national borders (OECD, 2019; UNESCO, 2021). Such a treaty would need to address both capability development and deployment practices, incorporating verification mechanisms and shared safety research agendas. Drawing lessons from existing regimes governing dual-use technologies, the treaty would seek to establish common baselines of safety and transparency while remaining adaptable to rapid technical change. Given the tendency of overlapping international regimes to fragment into a “regime complex” of only loosely coordinated institutions when no single body has full authority over a transnational risk (Keohane & Victor, 2011), particular attention should be paid to the treaty’s interoperability with existing standard-setting bodies (Cihon, 2019) and with regional and sectoral AI governance initiatives, so that the resulting architecture reinforces, rather than duplicates or contradicts, this wider institutional landscape.

The second is Technical Safety Research: the allocation of substantial resources to address the alignment problem, ensuring that advanced AI systems function within safe and reliable boundaries that remain consistent with human values (Amodei et al., 2016; Russell, 2019). Public funding for safety research is particularly important given the commercial incentives that currently prioritise capability advancement over robustness and interpretability. Research agendas should prioritise problems of scalable oversight, reward specification, and distributional robustness, while remaining attentive to the broader socio-technical context in which systems are deployed. Framing safety research purely as a race against rival institutions or states risks becoming counterproductive, since the rhetorical construction of an AI arms race can itself pressure researchers and funders towards premature deployment at the expense of the very safety work such a treaty is intended to

protect (Cave & ÓhÉigearthaigh, 2018); a credible governance architecture must therefore treat cooperative safety research as a global public good rather than as a by-product of competitive advantage.

6.2 Societal mechanisms: digital literacy, democratic dialogue, and distributive legitimacy

The third is Digital Literacy and Democratic Dialogue: the equipping of civil society with the tools to critically evaluate the implications of AI, ensuring that technology serves collective welfare rather than corporate interests or state domination (Jobin et al., 2019). Inclusive public deliberation is essential for legitimising governance arrangements and for surfacing values and concerns that may be overlooked in technocratic or industry-dominated forums. Legal frameworks that guarantee transparency and public participation can help ensure that governance remains democratically accountable rather than captured by concentrated interests.

This societal dimension of governance cannot be separated from the distributive concerns raised in Section 5. If the economic transition associated with AI produces winners and losers unevenly distributed across regions, sectors, and demographic groups – as the displacement-versus-reinstatement dynamic described by Acemoglu and Restrepo (2019) suggests it may – then digital literacy initiatives that are not accompanied by meaningful redistributive and participatory mechanisms risk becoming a form of responsabilisation, in which citizens are asked to adapt individually to structural transformations over which they have little collective influence. Effective democratic dialogue on AI governance therefore requires not only technical education, but also accessible channels through which affected communities, including those in low- and middle-income countries whose digital infrastructures and regulatory capacities differ substantially from those of the jurisdictions where most frontier AI research currently takes place, can shape the substantive content of the rules that will govern systems deployed in their societies.

7 Final considerations

This investigation explored the impact of Artificial Intelligence, framing it not merely as a technical innovation but as a revolutionary milestone for human civilisation and the contemporary legal order. By examining the spectrum between utopian potential and existential hazard, the study demonstrated that humanity's future in an AI-driven world is not predetermined by technology itself, but is shaped by present-day political, ethical, and legal choices.

The study confirmed its central hypothesis: the influence of AI is inherently tied to the equilibrium between the rapid advancement of technological capabilities and the strength of global governance systems. The findings emphasise that without effective value alignment mechanisms and regulatory frameworks that transcend national borders, humanity risks confronting severe challenges – including deepening inequality, the erosion of individual autonomy, and potential misalignment with superintelligent systems (Bostrom, 2014; Russell, 2019). Existing legal frameworks remain inadequate to the scale and novelty of the challenges posed by algorithmic autonomy; anticipatory, internationally coordinated regulation is therefore indispensable.

The analysis concludes that AI development demands a reimagined social contract, one in which innovation is guided by the precautionary principle and a commitment to collective well-being and human rights. Governance must serve as the guiding force for the ethical coevolution of technology, ensuring that AI enhances human dignity rather than undermines it. In navigating this era, society must strike a balance between uncritical optimism and undue pessimism with respect to emerging technologies, particularly AI. The capacity to govern AI effectively will serve as a critical test of the international system's ability to manage planetary-scale technological risks in the decades ahead.

A recurring theme across the scenarios and structural challenges examined in this article is that the distribution of AI's benefits and harms is not incidental to its technical trajectory but is actively shaped by who participates in setting the rules that govern it. This is true of standard-setting bodies, of international treaties, and of the civic and educational institutions through which societies deliberate about technology. Future research should therefore give particular attention to the institutional design of inclusive, enforceable, and adaptive governance mechanisms – especially those capable of incorporating the perspectives of regions currently under-represented in global AI policymaking – as a complement to the technical safety research on which the alignment problem itself depends.

Ultimately, the defining challenge of the twenty-first century lies not in the capacity to build intelligent machines, but in the wisdom to ensure that these systems operate in service of shared human values and an interconnected global community. The core dilemma does not reside in AI itself, but in the actions and motivations of humanity, shaped by geopolitical rivalry, economic incentives, and the pursuit of dominance. The choices made in the coming years will determine whether AI becomes a force for collective flourishing or a source of irreversible disruption.

Declarations

Data Availability: This article does not analyse or generate any new datasets. It is based on the review and synthesis of previously published academic literature, policy documents, and publicly available institutional reports, all of which are cited within the References section.

Ethical Approval: This article does not contain any studies with human participants performed by any of the authors.

Informed Consent: This article does not contain any studies with human participants performed by any of the authors.

References

1. Acemoglu D, Restrepo P (2019) Automation and new tasks: how technology displaces and reinstates labor. *J Econ Perspect* 33(2):3–30. <https://doi.org/10.1257/jep.33.2.3>
2. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D (2016) Concrete problems in AI safety. *arXiv*. <https://doi.org/10.48550/arXiv.1606.06565>
3. Bender EM, Gebru T, McMillan-Major A, Shmitchell S (2021) On the dangers of stochastic parrots: can language models be too big? In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Association for Computing Machinery, New York, p 610–623. <https://doi.org/10.1145/3442188.3445922>
4. Benjamin R (2019) *Race after technology. abolitionist tools for the new Jim code*. Polity, Cambridge
5. Bostrom N (2014) *Superintelligence: paths, dangers, strategies*. Oxford University Press, Oxford
6. Brynjolfsson E, McAfee A (2014) *The second machine age: work, progress, and prosperity in a time of brilliant technologies*. W. W. Norton & Company, New York
7. Cave S, ÓhÉigeartaigh SS (2018) An AI race for strategic advantage: rhetoric and risks. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. Association for Computing Machinery, New York, p 36–40. <https://doi.org/10.1145/3278721.3278780>
8. Cihon P (2019) *Standards for AI governance: international standards to enable global coordination in AI research & development*. Future of Humanity Institute, University of Oxford, Oxford
9. Dafoe A (2018) *AI governance: a research agenda*. Centre for the Governance of AI, Future of Humanity Institute. University of Oxford, Oxford
10. Eubanks V (2018) *Automating inequality: how high-tech tools profile, police, and punish the poor*. St. Martin's, New York
11. Fjeld J, Achten N, Hilligoss H, Nagy A, Srikumar M (2020) *Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI*. Berkman Klein Center Research Publication No. 2020-1. Harvard University, Cambridge, MA
12. Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C, Madelin R, Pagallo U, Rossi F, Schafer B, Valcke P, Vayena E (2018) *AI4People—an ethical framework for a good AI society*:

- opportunities, risks, principles, and recommendations. *Minds Mach* 28(4):689–707.
<https://doi.org/10.1007/s11023-018-9482-5>
13. Frey CB, Osborne MA (2017) The future of employment: how susceptible are jobs to computerisation? *Technol Forecast Soc Change* 114:254–280.
<https://doi.org/10.1016/j.techfore.2016.08.014>
 14. Goodfellow I, Bengio Y, Courville A (2016) *Deep learning*. MIT Press, Cambridge, MA
 15. Hagendorff T (2020) The ethics of AI ethics: an evaluation of guidelines. *Minds Mach* 30(1):99–120.
<https://doi.org/10.1007/s11023-020-09517-8>
 16. Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. *Nat Mach Intell* 1:389–399. <https://doi.org/10.1038/s42256-019-0088-2>
 17. Keohane RO, Victor DG (2011) The regime complex for climate change. *Perspect Polit* 9(1):7–23.
<https://doi.org/10.1017/S1537592710004068>
 18. Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. *Nat Mach Intell* 1(11):501–507.
<https://doi.org/10.1038/s42256-019-0114-4>
 19. Noble SU (2018) *Algorithms of oppression: how search engines reinforce racism*. New York University, New York
 20. O’Neil C (2016) *Weapons of math destruction: how big data increases inequality and threatens democracy*. Crown, New York
 21. OECD (2019) *Scoping the OECD AI principles: initial implementation actions for governments*. OECD Publishing, Paris
 22. Ord T (2020) *The precipice: existential risk and the future of humanity*. Bloomsbury, London
 23. Rahwan I, Cebrian M, Obradovich N, Bongard J, Bonnefon J-F, Breazeal C, Crandall JW, Christakis NA, Couzin ID, Jackson MO, Jennings NR, Kamar E, Kloumann IM, Larochelle H, Lazer D, McElreath R, Mislove A, Parkes DC, Pentland A, Roberts ME, Shariff A, Tenenbaum JB, Wellman M (2019) Machine behaviour. *Nature* 568(7753):477–486. <https://doi.org/10.1038/s41586-019-1138-y>
 24. Reisman D, Schultz J, Crawford K, Whittaker M (2018) *Algorithmic impact assessments: a practical framework for public agency accountability*. AI Now Institute, New York
 25. Russell S (2019) *Human compatible: artificial intelligence and the problem of control*. Viking, New York
 26. Scharre P (2018) *Army of none: autonomous weapons and the future of war*. W. W. Norton & Company, New York
 27. Schwartz R, Dodge J, Smith NA, Etzioni O (2020) Green AI. *Commun ACM* 63(12):54–63.
<https://doi.org/10.1145/3381831>
 28. Strubell E, Ganesh A, McCallum A (2019) Energy and policy considerations for deep learning in NLP. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, p 3645–3650.
<https://doi.org/10.18653/v1/P19-1355>

29. Susskind D (2020) A world without work: technology, automation, and how we should respond. Metropolitan Books, New York
30. UNESCO (2021) Recommendation on the ethics of artificial intelligence. UNESCO, Paris
31. UNICRI INTERPOL (2021) Towards responsible AI innovation: second report on artificial intelligence for law enforcement. United Nations Interregional Crime and Justice Research Institute, Turin
32. World Economic Forum (2020) The future of jobs report 2020. World Economic Forum, Geneva
33. Zuboff S (2019) The age of surveillance capitalism: the fight for a human future at the new frontier of power. PublicAffairs, New York