GOVERNANCE OF ARTIFICIAL INTELLIGENCE AND ITS IMPACT ON THE TRANSFORMATION OF SOCIETIES

Prof. Dr. Mehmet Yildiz
Sabanci University

Attribution: Yildiz, M. (2026). Governance of artificial intelligence and its impact on the transformation of societies. In H. Arslan, A. Taşdelen, E. Kuzucu Hıdır, & O. A. Çetin (Eds.), *Artificial intelligence and national technology reflections* (pp. 457–479). Turkish Academy of Sciences. https://doi.org/10.53478/TUBA.978-625-6110-86-1.ch19

10.53478/TUBA.978-625-6110-86-1.ch19

GOVERNANCE OF ARTIFICIAL INTELLIGENCE AND ITS IMPACT ON THE TRANSFORMATION OF SOCIETIES

Prof. Dr. Mehmet Yildiz
Sabanci University

Abstract

Few technological shifts in modern history have arrived as quickly, or landed across as many domains at once, as the rise of artificial intelligence (AI). In a decade, AI has moved from the domain of computer science into the operational core of healthcare, finance, scientific research, education, and government. Machine learning, deep learning, and, most recently, large foundation models have driven this expansion faster than most policy and governance frameworks are formed to handle. The result is the emergence of a technology which is enabling and disruptive at the same time. AI offers remarkable possibilities for societies, including scientific acceleration, productivity growth, and universal access to knowledge. However, it comes with risks such as algorithmic discrimination, erosion of privacy, labor displacement, high-volume misinformation, and the accumulation of critical capabilities in a handful of institutions and nations. The workforce dimension alone demands serious attention. The World Economic Forum projects that approximately 92 million existing roles will be displaced globally, whereas around 170 million new ones will emerge, hence yielding a net gain of roughly 78 million jobs, but only under conditions of successful and large-scale reskilling. By 2040, half or more of today's jobs may be automated or transformed, and some sectors may experience that figure approach 80% by 2050. These are not distant science fiction scenarios. They points to changes already underway. This paper examines what those changes mean in practice, and focuses on the domains where the effects of AI are most substantial, such as higher education, scientific research, labor markets, governance, and environmental sustainability, among others. The argument proposed here is that AI cannot be adequately understood as simply another productivity tool. AI is becoming an integral part of modern civilization through reshaping how knowledge is created and redefining what human work and learning mean. Whether the reshaping and redefining in questions serves human flourishing or undermines it will depend less on the technology itself than on the political, institutional, and ethical decisions societies make in the years immediately ahead.

Keywords

Artificial Intelligence, Higher Education, Workforce Transformation, AI Governance, Human Cognition

Introduction

When writing about artificial intelligence, there is an immediate temptation to reach immediately for superlatives. This is understandable as the penetration of AI across sectors is unprecedented. The activities being affected by AI range from radiological diagnosis, research on protein structure prediction to legal and societal matters. These areas are broad enough to justify praising adjectives and terminologies for AI. Yet, the more analytically useful starting point is a simpler observation. AI systems are doing things which, until very recently, have required trained human judgment. That shift carries implications that go well beyond the economics of automation (Brynjolfsson & McAfee, 2014).

What distinguishes today from earlier waves of technological change involves both the sophistication of the systems and the nature of the cognitive territory we are in. The mechanization of physical labor during the Industrial Revolution displaced workers from mines and factories. Nevertheless, it did not threaten the activities defining professional and intellectual life. Nonetheless, AI does. It writes, reasons, interprets, creates, and advises, albeit imperfectly, yet increasingly well enough to alter the economics of many knowledge intensive occupations. Understanding this distinction is, I would argue, the essential precondition for thinking clearly about the social consequences of AI (Brynjolfsson & McAfee, 2014).

A second important observation is associated with the rate of change with respect to the adaptation pace of institutions. Electrification, industrialization, and the digitalization of the 1990s all transformed economies profoundly, but they did so over timescales which allowed legal systems, educational institutions, and labor markets to evolve in rough synchrony. AI is developing faster. Regulatory frameworks, curricula, and professional training designed for a different technological environment are now clearly struggling. Dafoe (2018) has described this condition as a form of governance lag which is one of the most serious risks of the current period.

The technical elements enabling all of this, including exponential growth in computational power, vast training datasets, advances in semiconductor design, and the architectural breakthroughs, are well documented in literature. What this paper is primarily concerned with are the consequences of how universities teach, how science is conducted, who works and on what terms, how democracies govern information, and whether the environmental costs of AI development can be made compatible with the sustainability commitments countries have nominally adopted (Bommasani et al., 2021; Russell & Norvig, 2020).

1. Historical Evolution and Technological Foundations

1.1. From Symbolic Reasoning to Statistical Learning

The history of AI involves repeated unfulfilled promises followed by unexpected resurgence. Early researchers in the 1950s and 1960s, including Turing, McCarthy, Minsky, among others, supported an essentially optimistic view. That is to say, human intelligence could be captured in explicit logical rules, and machines given those rules could exhibit something which can be considered as intelligence. This expectation largely failed. Rule based systems worked well in limited, well-defined domains, but proved fragile when challenged with ambiguity and context dependence, which characterize most real-world problems (Turing, 1950).

The expert systems of the 1970s and 1980s were employed in bounded domains such as medical diagnosis, geological surveying, and financial analysis, where structured knowledge could be encoded in decision trees. They provided obvious commercial values but remained fundamentally limited. This was due to the fact that these systems could not learn, could not generalize, and required enormous human effort for maintenances as the domains they covered evolved.

The shift to machine learning in the 1990s and early 2000s changed the architecture of the problem entirely. Rather than encoding the knowledge explicitly, researchers have built systems capable of inferring statistical relationships from data. The deep learning revolution followed in the 2010s, powered by GPU accelerated training and the large-scale visual recognition abilities showed that multilayer neural networks could achieve performance on perceptual tasks, including image recognition, speech transcription, and language translation, which were before regarded beyond reach (Goodfellow et al., 2016; Krizhevsky et al., 2012).

1.2. Foundation Models and the Current Inflection Point

The most recent transition to foundation models and generative AI represents something qualitatively different from before. Earlier machine learning systems were trained for specific tasks: a system trained to detect tumors in chest scans could do that, and little else. Large language models (LLM) and multimodal systems trained on diverse internet scale datasets can generate text, images, code, and audio; answer questions across domains; reason through novel problems; and be adapted to specialized applications with relatively modest additional training (Brown et al., 2020). The research literature describes this capacity for broad generalization as emergent, referring to capabilities that appear at scale without being explicitly trained for and that were not predictable from the behavior of smaller models (Wei et al., 2022).

This generality is what makes the current generation of AI systems different, and what makes the governance questions surrounding them so difficult. A narrow AI system deployed in a specific context can be evaluated, regulated, and held accountable in relatively straightforward ways. A general-purpose system that can be adapted to thousands of different applications by thousands of different actors is a categorically harder regulatory problem, and that is the situation we now face.

2. Global AI Competition, National Strategies, and Geopolitical Dynamics

2.1. The Three-Way Contest

The geopolitics of AI are mainly formed around a competition among the United States, China, and the European Union. Each actor countries follows a recognizably distinct strategy which reflects its domestic institutional arrangements and its broader perspectives of what AI is ultimately for.

The United States has relied primarily on private sector dynamism, with a small number of very large technology companies, including OpenAI, Google DeepMind, Microsoft, Meta, Anthropic, and NVIDIA, collectively investing at a scale that no government program has matched. The benefits are real, but they do not reach everyone equally. These benefits flow mostly to the companies involved, their investors, and their research partners, not to the public at large.

The AI model of China is more hands-on and state-driven. National AI development plans, substantial state investment in semiconductor infrastructure, and the deployment of AI within surveillance and social management systems indicates the AI model in which the developed technologies serve governmental as well as commercial goals. The ethical tensions this creates, particularly around privacy, civil liberties, and the use of AI in law enforcement, are ones that the Chinese governance framework handles quite differently from liberal democracies.

The European Union has chosen a different approach. The governance dimension of AI is not merely a constraint on innovation but a potential competitive advantage. The EU AI Act of 2024, which classifies AI applications by risk level and imposes corresponding obligations on developers and deployers, represents the most comprehensive attempt so far to establish a legal framework for AI. Whether it will achieve its ambitions or primarily succeed in disadvantaging European firms relative to less regulated competitors remains an open question (European Commission, 2024).

Other countries, including Canada, Singapore, Japan, South Korea, and the United Kingdom, have developed national strategies that vary in emphasis but tend to converge around the same core elements: compute infrastructure, data governance,

talent development, and collaboration between academia and industry. What this convergence suggests is that AI capability is not reducible to algorithmic sophistication alone. It depends on an entire institutional ecosystem, and countries that lack strong universities, functioning regulatory frameworks, and deep capital markets face structural disadvantages that better algorithms cannot fully compensate.

Higher education sits inside this three-way contest in ways that mirror each actor's broader posture. In the United States, universities function largely as autonomous partners to industry. Top institutions such as Stanford and MIT, among others supply talent and fundamental research to the same companies driving commercial AI development, with relatively little central coordination of curricula or research priorities from the federal government. This produces world-leading research output but leaves questions of equitable access to AI literacy and training largely to individual institutions and states. China, by contrast, treats higher education as an instrument of national AI strategy. Universities are directed to align degree programs, research agendas, and talent pipelines with state industrial policy, and AI literacy has been pushed into curricula from the secondary level upward as part of a coordinated, top-down effort to build domestic capability. The European Union pursues a middle path, using instruments such as the Digital Europe Programme and European university alliances to promote harmonized AI curricula and cross-border research collaboration, while its regulatory posture, discussed further below, also shapes what universities can teach about deployment and governance. The comparison suggests that a country's higher education strategy is not a peripheral input to AI competitiveness but one of its central drivers, and that the three actors are, in effect, running three different experiments in how to organize that relationship.

2.2. Semiconductors, Defense, and Strategic Risk

Beneath the AI competition lies a more fundamental competition over semiconductor manufacturing capability. Advanced AI training and inference require chips, primarily GPUs and specialized accelerators, which can only be manufactured by a handful of companies, most of them located in Taiwan, South Korea, the Netherlands, and the United States. The strategic implications of this concentration became apparent when the United States imposed export controls on advanced chips sold to China in 2022 and 2023. This triggered responses which have reshaped global supply chains and investment patterns.

The integration of AI into military systems adds another dimension of concern. Autonomous drones, AI assisted targeting, algorithmic surveillance, and the use of AI in cyber operations are all areas of active development across multiple militaries. The ethical and strategic risks, among them reduced decision timelines, accountability gaps when autonomous systems cause harm, and escalation dynamics that human

judgment might have interrupted, are ones that existing international law is poorly equipped to address. This gap between technological capability and governance infrastructure may prove to be the most consequential failure of the current period (Scharre, 2018).

3. AI Transformation in Higher Education, Scientific Research, and Human Cognition

3.1. What Universities Are Actually Facing

Higher education institutions occupy an unusual position in the AI landscape. They are simultaneously major sites of AI research, primary targets of the disruption AI brings, and the institutions most responsible for preparing the next generation to navigate both. This triple role creates tensions that have not yet been honestly confronted in most institutional responses to AI.

The pedagogical opportunities are real. Adaptive learning systems that respond to individual student performance patterns have shown measurable improvements in outcomes in controlled settings. Intelligent tutoring systems can provide the kind of immediate, personalized feedback that human instructors cannot realistically offer at scale, and controlled studies have found them to produce learning gains comparable in some contexts to one-to-one human tutoring (Kestin et al., 2025). Virtual laboratories and simulation environments make certain forms of practical training accessible to students who would otherwise lack them. These are genuine advances, and dismissing them as threats to traditional pedagogy misses something important (Luckin, 2018).

But the threats are equally real, and in some respects more urgent. The widespread adoption of generative AI for student writing has created an academic integrity crisis that existing honor codes and plagiarism detection tools were not designed to handle. More fundamentally, there is a serious question about whether students who routinely outsource analytical and compositional work to AI systems are developing the intellectual capacities that higher education is supposed to cultivate. The evidence is mixed, but the concern that passive reliance on AI generated outputs may attenuate the effortful processing that produces genuine understanding is grounded in what we know about how learning works (Kasneci et al., 2023).

This puts universities in a uniquely difficult position. They cannot simply prohibit AI use since the tools are too pervasive and the legitimate applications too valuable. But uncritical acceptance risks turning higher education into a credentialing exercise in which the credential increasingly measures competence at directing AI systems rather than independent intellectual capability. Finding a coherent institutional response to this tension is, I would argue, the most important governance challenge currently facing higher education.

3.2. AI as a Scientific Instrument

The transformation of scientific research by AI is already well advanced in several fields, and the trajectory is clear. In genomics, climate science, drug discovery, materials research, and astrophysics, AI is no longer a peripheral analytical aid. It is becoming central to how science is actually done.

The reason is partly about data volume. Modern scientific instruments generate information at scales that would have been inconceivable two decades ago. Particle accelerators, genome sequencers, astronomical observatories, and climate monitoring networks together produce petabytes of data annually. Human researchers simply cannot analyze that volume unaided. As such, AI systems are not optional enhancements but necessary infrastructure.

But the more interesting development is AI's growing role in hypothesis generation and experimental design, functions that were previously regarded as requiring specifically human creativity. AlphaFold's prediction of protein structures with near experimental accuracy from amino acid sequences alone is the most celebrated example. Namely, it does not merely accelerate work that researchers have been already doing, but, it enables work that was previously impossible (Jumper et al., 2021). Similar dynamics are playing out in drug repurposing, materials discovery, and climate modeling (Rolnick et al., 2022).

There are real concerns alongside these advances. The use of AI in manuscript preparation, literature synthesis, and data analysis has raised questions about authorship attribution and scientific accountability that existing publication norms are struggling to resolve. More broadly, if AI systems become central to hypothesis generation, the question of whether we understand why they generate hypotheses they do, and what that means for scientific knowledge, becomes pressing in ways that the philosophy of science has barely begun to address.

3.3. What AI Does to Human Thinking

The effect of AI on human cognition is, in my view, the most under-examined dimension of the current transformation. This is partly because it is methodologically difficult to study, and partly because the timeframe over which the most significant effects will manifest is longer than the news cycle or the typical research grant.

The historical analogies are instructive but imperfect. Writing did reduce reliance on memory, as Carr (2010) and others have noted, but it also enabled forms of intellectual complexity that oral culture could not sustain. The printing press did change how people related to authoritative knowledge, but it also created conditions for scientific revolutions. Calculators reduced the need to perform arithmetic by hand, yet freed mathematicians for work that actually required mathematical thinking. Each technology displaced some cognitive activity while enabling other, often richer, forms of it.

The question worth taking seriously is whether AI is likely to follow the same pattern, or whether there is something distinctive about the cognitive activities AI is entering, activities more central to human intellectual development, that makes the displacement more problematic. The concern, explored at length by Carr in the context of internet use and now being raised with fresh urgency about AI, is that the effortful process of producing thought, including writing a draft, working through proof, and synthesizing literature, is not merely instrumental to a product. It is constitutive of understanding. If that is right, then AI systems that remove the friction may also, under conditions of passive use, remove the learning (Carr, 2010).

At the same time, AI extends certain forms of human analytical capacity in ways that are unambiguously valuable, thereby enabling researchers to find patterns in datasets too large to analyze manually, allowing practitioners to access diagnostic or legal reasoning that would otherwise require specialists, and supporting communication across linguistic boundaries at a scale no prior technology permitted. Clark's argument that humans have always been cognitive hybrid creatures, extending their mental reach through tools, provides a useful frame here. To be specific, the question is not whether AI constitutes a cognitive extension, but whether the particular form of extension it offers supports or undermines the underlying capacities it displaces (Clark, 2003). That is a harder design problem than either AI enthusiasts or AI skeptics tend to acknowledge.

It is worth being more precise about which cognitive abilities are actually at stake, since "the mind" is not a single faculty that AI either preserves or erodes. Cognitive science distinguishes at minimum between working memory, the capacity to hold and manipulate information over short intervals; long-term encoding, the consolidation of that information into durable knowledge; and executive function, the capacity to plan, evaluate, and revise one's own reasoning. The effortful activities this paper has been discussing, drafting, proving, synthesizing, engage all three simultaneously: working memory manipulates the problem, encoding builds the schema that later becomes expertise, and executive function monitors whether the emerging argument actually holds together. When an AI system supplies a finished draft or a completed derivation, it is not simply saving time; it is removing the specific cognitive operations, rehearsal, error detection, self-correction, through which working memory content is transferred into durable long-term knowledge and executive control is trained. This is why the concern about AI and cognition cannot be resolved by pointing to calculators or search engines as reassuring precedents: those tools offloaded narrow, well-defined operations while leaving the surrounding reasoning to the human user, whereas generative AI can offload the reasoning itself. Which specific abilities are protected or eroded will therefore depend on how a given tool is used, whether as a check on reasoning the person has already done, or as a substitute

for doing that reasoning at all, a distinction that current usage patterns rarely make explicit and that future research on AI and cognition will need to measure directly rather than infer from aggregate outcomes.

4. Productivity, Automation, and Workforce Transformation

4.1. What the Numbers Actually Say

The economic projections discussed in this section concern automation, the substitution or restructuring of specific job tasks by AI systems operating under human direction, rather than autonomy, which can be defined as the capacity of a system to set its own goals and act without human oversight. The distinction matters because the two raise largely separate questions: automation is primarily a labor-market and productivity question, addressed in this section, whereas autonomy is primarily a control and accountability question, addressed later in this paper's discussion of military and safety-critical systems (Section 3.2). The economic projections associated with automation driven by AI have become a familiar feature of public debate, often cited in ways that strip away the substantial uncertainty they carry. It is worth being precise about what is and is not known.

The World Economic Forum's 2025 Future of Jobs report projects net global job creation, roughly 78 million more jobs created than destroyed, but under a condition that is itself deeply uncertain: successful large-scale reskilling. McKinsey's analysis suggests that AI could automate around 30% of current work hours by 2030, with productivity gains adding 0.5 to 0.9 percentage points to annual global economic growth (McKinsey Global Institute, 2023; World Economic Forum, 2025). These are plausible central estimates, but the confidence intervals around them are wide, and the distributional question of who captures the productivity gains and who bears the displacement costs is largely orthogonal to the aggregate numbers.

What is clearer is the pattern of exposure across occupational categories. Roles involving routine information processing, including clerical work, administrative support, customer service, and data entry, face high automation probability. Roles requiring physical dexterity in unpredictable environments, complex interpersonal interaction, or genuine creative and strategic judgment face lower short-term exposure, though this boundary is eroding as AI capabilities expand. The asymmetry matters enormously for policy such that the workers most at risk of displacement are typically those with the fewest resources for independent reskilling (McKinsey Global Institute, 2023).

4.2. The Structural Transformation Problem

Looking further out, to 2040 and beyond, the projections become more dramatic and more speculative. Between 50% and 60% of current jobs may be substantially automated or transformed by 2040. In some sectors, the figure may exceed 80% by

2050. If these trajectories materialize, the scale of required labor market transition would exceed anything in modern economic history, including the deindustrialization of Western economies in the 1970s and 1980s (McKinsey Global Institute, 2023; World Economic Forum, 2025).

The transition challenge can be usefully decomposed into three distinct processes. First, what might be called rescaling: certain sectors shrink substantially while new sectors with intensive AI use expand, thus requiring workers to move not just between employers but between entire economic domains. Second, upskilling: even workers who remain in their current sectors will need to acquire new competencies as the human tasks within those roles evolve. Third, transitioning: for workers whose entire occupational category disappears or contracts severely, large scale retraining into new fields is required. Each of these processes requires a different policy response, and all three are simultaneously insufficient in most economies (OECD, 2023).

4.3. The Concentration Problem

One of the most important and least adequately discussed economic dimensions of AI is the concentration dynamic it tends to generate. AI capability scales with compute, data, and talent, all of which exhibit strong returns to scale. The result is that a small number of very large firms, primarily based in the United States and, to a lesser extent, China, have accumulated advantages in AI development that are proving difficult for smaller competitors or smaller economies to close. This is not merely a market competition concern. It has implications for national economic sovereignty and for the distribution of the wealth that AI generates (Acemoglu & Restrepo, 2018).

Against this, there is a genuine democratizing tendency within generative AI. The cost of building software applications, producing written content, and accessing sophisticated analytical tools has fallen sharply for smaller organizations and individuals. This may partially offset concentration at the infrastructure layer by distributing capability enabled by AI more broadly at the application layer. Whether the net effect is more or less concentration of economic power is an empirical question that will likely play out differently across different sectors and national contexts.

Longer-term distributional questions about whether the productivity gains AI generates will flow primarily to capital or be more broadly shared have prompted serious proposals for structural policy responses: Universal Basic Income, reduced standard working hours, redistribution mechanisms tied to AI productivity dividends, and various forms of public or cooperative ownership of AI infrastructure. Most economists regard these proposals as politically premature. The more urgent question may be whether existing tax and labor market frameworks are adequate to prevent a significant worsening of income inequality as AI adoption accelerates (Acemoglu & Restrepo, 2018).

5. AI Applications in Healthcare and Creative Industries

5.1. Medicine as a Test Case

Healthcare is worth examining in some depth as a domain where both the promise and the governance challenges of AI are unusually clear. The clinical applications are truly impressive and, in some cases, already affecting patient outcomes.

In radiology, AI systems trained on large labeled datasets have achieved sensitivity and specificity rates for detecting certain abnormalities, including lung nodules, diabetic retinopathy, and breast cancer on mammography, that are comparable to, and in some settings exceed, those of experienced clinicians. In genomics, the ability to analyze DNA sequences at scale has made personalized medicine, which is treatment tailored to the patient's specific molecular profile, practically achievable in ways it was not a decade ago. AlphaFold's prediction of protein structures has done for structural biology what the Human Genome Project did for genetics. It has unlocked a research agenda that would otherwise have taken decades (Jumper et al., 2021; Rajpurkar et al., 2022).

Wearable devices and continuous monitoring systems are enabling a further shift from episodic healthcare (diagnosis and treatment when a patient presents with symptoms) toward continuous monitoring which can identify early warning signals before needing a clinical service. For example, passive detection of atrial fibrillation using commercially available smartwatches indicated how continuous monitoring capabilities is already transitioning from laboratory to practice. Essentially, the boundaries between consumer electronics and medical devices have started disappearing (Tison et al., 2018). For conditions responsive to early detection, this represents a qualitative improvement in care, not merely an incremental one.

However, the governance problems are still serious and not resolved. To elaborate, the study by Obermeyer et al. (2019) addressed that a widely used commercial algorithm systematically underestimated the health needs of black patients relative to white ones with equivalent clinical conditions. This is a form of bias due to the past care allocation. The source of this bias is not algorithmic error but the use of healthcare costs as a metric for health care needs, reflecting existing inequalities in access to care. This is a very good and representative example of a general problem. Specifically, AI systems trained on historically generated data learn from and therefore maintain the inequalities embedded in that history.

Legal accountabilities arising from adverse clinical outcomes associated with AI systems remain largely unsettled in most jurisdictions. The existing malpractice rules or standard frameworks for professional liability is centered on the individual clinician's duty of care. This framework is not able to address a situation where

clinical decisions are substantially shaped by an algorithmic tool that even its developers do not fully understand. The intersection of privacy law and cybersecurity in medical data creates an additional layer of governance complexity which standard regulatory frameworks can only partially address (Price & Cohen, 2019).

5.2. Creativity and the Question of Originality

The entry of AI into creative domains brings questions which are simultaneously economic, legal, and philosophical, and these different registers do not sit easily together.

The economic facts are clear enough. The generative AI can produce images, music, text, and code which are commercially usable, in large volumes, at very low cost. This does and will continue to displace demand for human creative labor in certain segments, including stock imagery, certain categories of commercial writing, and coding tasks. How far up the quality gradient this displacement extends is uncertain and is likely to vary considerably across creative domains.

The philosophical questions are harder. When an AI system generates an image in the style of a living artist, drawing on training data which included the artist's work, the concepts of originality, authorship, and creative labor become difficult to apply. Copyright law in most jurisdictions does not provide clear answers, and the pace of litigation has not yet resolved the questions courts will eventually need to decide.

There is also a cultural question worth taking seriously. Namely, whether environments in which high quality creative content is available at zero marginal cost will support or undermine the human creative practice which generated the training data in the first place. Some argue that cheaper AI generated content will expand the overall market for creative work. Others contend that it will devalue the skills and careers of human practitioners to the point of unsustainability. The evidence is too early and too partial to answer this question.

6. Ethics, Governance, and Regulatory Frameworks

6.1. The Bias Problem

Algorithmic bias is perhaps the most studied ethical challenges associated with AI, and the one where the gap between documented harm and effective policy response is most frustrating. The evidence that AI systems deployed in high stakes domains, including hiring, lending, criminal justice, and healthcare, routinely produce systematically different outcomes for different demographic groups is substantial and growing. The technical literature on fairness, accountability, and transparency in machine learning has advanced considerably, but the translation of that literature into meaningful operational practice has not experienced that progress (Jobin et al., 2019).

Part of the difficulty is conceptual. Fairness is not a single property but a family of properties which are, under most realistic conditions, mutually incompatible. A system cannot simultaneously equalize false positive rates across groups, equalize false negative rates across groups, and maintain calibration. These are mathematical constraints, not design failures (Barocas et al., 2023). Choosing among these criteria requires normative judgments that are inherently contestable, and that cannot be delegated to technical experts without misrepresenting their character.

Surveillance and privacy erosion represent a related but distinct set of concerns, and one worth connecting explicitly back to the bias problem just discussed. The connection is twofold. First, the surveillance infrastructure described below is itself the raw material from which many biased systems are built: facial recognition, behavioral profiling, and location or purchase data are exactly the inputs later fed into hiring, lending, and policing algorithms, so that whatever historical inequities are embedded in how that data was collected, who was watched, stopped, or flagged more often, propagate directly into the fairness problems discussed above. Second, surveillance and bias share a common victim profile: the populations subject to the most intensive algorithmic monitoring, including low-income communities and racial minorities, tend to be the same populations for whom the biased outcomes documented above are most severe. Surveillance is not merely a parallel harm to algorithmic bias; it is frequently the mechanism that generates and concentrates it. The combination of facial recognition, behavioral profiling, and the analysis of digital traces, including location data, purchase records, and social network connections, gives governments and large corporations access to information about individuals that previous surveillance regimes could not have assembled. Zuboff's (2019) analysis of what she calls surveillance capitalism describes this as a structural transformation of the relationship between citizens and institutions, not merely an incremental expansion of data collection.

6.2. Regulatory Architecture and Its Limits

The governance principles that have achieved broadest international consensus, among them beneficence, non-maleficence, transparency, accountability, explainability, and human oversight, are coherent as ethical commitments but remain loosely specified as operational requirements. The field of AI ethics has produced an abundance of principle statements and a relative shortage of enforcement mechanisms (Jobin et al., 2019). The EU AI Act represents the most serious attempt so far to translate principles into legally binding obligations, and its risk tiered architecture, distinguishing unacceptable, high, limited, and minimal risk applications, provides a more workable regulatory framework than the alternatives (European Commission, 2024). It will, however, face significant implementation challenges, and its extraterritorial reach remains contested.

Operational governance mechanisms, such as red teaming, algorithmic audits, model documentation, and institutional review processes, have gained traction within leading AI development organizations. These are valuable, but they are essentially voluntary and unevenly applied. The concepts of human in the loop and human on the loop offer a useful framework for thinking about where human judgment must be maintained, but they require specificity about what kinds of human oversight are meaningful. Doshi-Velez and Kim (2017) have argued that interpretability requirements are not merely procedural formalities but necessary conditions for accountability in high-stakes automated decisions, since a human who cannot understand the basis of a system's output cannot meaningfully evaluate or override it.

The deepest challenge of governance is jurisdictional. AI systems are developed by multinational corporations, deployed across national boundaries, and regulated by national legal systems that have limited tools for extraterritorial enforcement. This creates genuine risks of regulatory arbitrage, where companies locate development or deployment in jurisdictions with weaker standards, and makes international coordination not merely desirable but necessary (Dafoe, 2018; Floridi et al., 2018).

7. Environmental Sustainability and the Green AI Imperative

The environmental costs of AI are not widely appreciated, and the existing discourse tends to treat them as peripheral to the main concerns. This is a significant analytical error. The training of large foundation models consumes energy at a scale that is remarkable. Estimates suggest that training a single large language model can produce carbon emissions comparable to the lifetime emissions of several automobiles, though these figures vary substantially with the energy mix of the data center involved and have been improving as efficiency research matures (Patterson et al., 2021). Data center water consumption for cooling purposes, a less discussed but increasingly significant factor, is growing rapidly alongside energy demand (Schwartz et al., 2020).

This creates what is genuinely a paradox, not merely an apparent tension. AI is simultaneously one of the most promising tools available for addressing climate change, through improved climate modeling, renewable energy optimization, smart grid management, precision agriculture, and materials discovery for clean energy applications, and a significant and growing source of energy demand. The net environmental effect of AI deployment will depend heavily on the pace of decarbonization of electrical grids and on whether the efficiency gains AI enables in energy systems outpace the energy costs of AI operations themselves (Rolnick et al., 2022). This is an empirical question, not a rhetorical one, and the answer is not currently known.

The semiconductor supply chain adds a further layer of environmental concern. The production of advanced chips requires rare earth elements, substantial water consumption, and manufacturing processes which generate significant waste streams. The geographic concentration of chip manufacturing in a small number of regions, principally Taiwan, South Korea, and the Netherlands, means that expanding production capacity involves not just investment but extended supply chains with their own environmental footprints, causing concerns about sustainability that the industry has only recently begun to address systematically (Harrington et al., 2022).

The Green AI framework that has emerged within the research community emphasizes energy efficient architecture, carbon reporting for training runs, and the development of smaller but more capable models. Schwartz et al. (2020) have argued that the field needs to move beyond treating computational scale as the primary metric of AI progress, and that environmental efficiency should be incorporated into how AI systems are evaluated and compared. This argument seems right, and the practical challenge is getting funding structures and publication norms to reflect it.

8. AI, Democracy, Human Identity, and International Cooperation

8.1. The Integrity of Information Environments

Perhaps the most immediate threat that AI poses to democratic governance is not through labor markets or surveillance systems, but through the integrity of the information environment on which democratic deliberation depends. The capacity of generative AI systems to produce persuasive text, realistic images, and synthetic video at scale, and to target that content to specific audiences with algorithmic precision, creates conditions for influence operations of unprecedented sophistication.

Deepfake technologies capable of producing synthetic videos of public figures saying things they never said are real concern. Chesney and Citron (2019) identified the legal risks of this technology early. Essentially, they were concerned that AI would be able produce fake audio and video content nearly indistinguishable from authentic recordings. What was once considered as a potential risk has now become an immediate and real problem. Another problem is that even corrected misinformation leaves traces. As such, the exposure to deepfakes affect beliefs among audiences who are later told the content was fabricated (Vaccari & Chadwick, 2020).

Recommendation algorithms, though predating the current generation of generative AI, have already substantially altered how citizens encounter political information, with well documented effects on polarization and exposure to misinformation. Pariser's (2011) concept of the filter bubble has proven both prescient and somewhat too simple. The mechanisms are complex, the effects are heterogeneous across

platforms and populations, and the policy responses remain contested. What is clear is that algorithmic curation of political information at scale creates dynamics which differ in kind, not merely degree, from those of earlier broadcast media environments.

Governance responses under development, including content watermarking, provenance tracking systems, platform transparency requirements, and algorithmic accountability measures, represent serious attempts to address these problems. But, they face a fundamental asymmetry. The tools for generating synthetic content are advancing faster than the tools for detecting it. Additionally, the regulatory frameworks for governing digital content remain fragmented across jurisdictions in ways that limit their effectiveness. Balancing these responses with the protection of legitimate free expression remains difficult, and there is no technical solution that resolves the normative conflict.

8.2. Human Identity under Technological Pressure

The philosophical questions raised by AI about human identity and cognitive distinctiveness are genuine, if sometimes discussed in more dramatic terms than the evidence currently supports. Humans have historically defined themselves partly in terms of intellectual capacities that AI is now beginning to approximate, not because there is anything metaphysically special about those capacities, but because they have organized professional life, social status, and self-conception in consequential ways.

When AI systems produce outputs, including legal briefs, diagnoses, creative works, and strategic analyses, which are difficult to distinguish from those produced by trained human professionals, the social and psychological implications go beyond the labor market. They raise questions about what kind of preparation and cultivation is worth investing in, what the relationship between effort and expertise becomes when AI can reproduce some of its products without the effort, and how institutions that have historically certified competence through the demonstration of skilled performance should adapt. These are not issues with clear solutions or answers, but they require ongoing social negotiation rather than technical resolution.

The worry about cognitive dependency, that continuous reliance on AI for tasks that were previously performed by human judgment will gradually erode the capacities that make that judgment possible, deserves serious attention even if it cannot currently be empirically confirmed. Clark's argument that humans have always been cognitive hybrid creatures, extending their mental reach through tools ranging from writing to calculators to search engines, provides a useful frame: the question is not whether AI constitutes a cognitive extension but whether the particular form of extension it offers supports or undermines the underlying capacities it displaces (Clark, 2003).

8.3. The Case for International Governance

The problems discussed throughout this paper, including bias, privacy erosion, labor displacement, misinformation, environmental costs, and autonomous weapons, share an important characteristic. They do not respect national borders and cannot be adequately addressed by any single jurisdiction acting alone. This creates a strong case for international AI governance frameworks, even if the practical politics of building such frameworks are formidable.

The existing multilateral architecture, encompassing the OECD AI Principles, UNESCO's Recommendation on the Ethics of AI, and the work of the UN's AI Advisory Body, represents a start, but these frameworks are largely of recommendatory nature since they lack enforcement mechanisms and their adoption is uneven. The challenge is that the nations most capable of shaping effective international frameworks, namely the United States, China, and the EU, among others have divergent interests and governance philosophies that make consensus genuinely difficult to achieve (Floridi et al., 2018).

The risks of failure are asymmetric. In a world of fragmented AI governance, the standards that prevail will tend to be those of the most permissive jurisdictions, creating a race to the bottom in regulatory standards that the strongest actors have limited incentives to resist. International coordination is not just normatively desirable; it is practically necessary if governance is to be effective rather than merely symbolic (Bengio et al., 2024).

Conclusion

The aim of this paper is to examine transformative effects of AI across multiple domains without reducing them to a single story. The single-story versions are readily available in literature in the context of AI as liberation technology, facilitating the accessibility to knowledge and accelerating science or AI as threat, displacing workers, amplifying bias, and concentrating power. Both contain truth, and both are insufficient.

What the cross-domain analysis reveals is competing results. Namely, the same systems that accelerate scientific discovery create new vulnerabilities for scientific integrity. The same capabilities that enable personalized education bring up questions about the usage of independent intellectual capacity. The same productivity gains which may increase aggregate wealth tend to concentrate that wealth in ways that existing distributional frameworks are not designed to counteract. The same tools that can optimize renewable energy systems consume substantial amounts of the energy we are trying to optimize. These competing results are not accidental

features of particular AI applications. Rather, they reflect something structural about the technology and how it is currently being developed and deployed.

The workforce projections, 92 million jobs displaced, 170 million created, a net gain of 78 million under successful reskilling, are often cited as grounds for reassurance. They should instead be read as a specification of the policy challenge. The reskilling scenario that produces the net gain is not a default outcome, it is a goal that requires sustained and coordinated investment in education, labor market infrastructure, and social protection systems. The historical record of how societies have managed previous technological transitions, sometimes well, often poorly, does not justify complacency.

On the governance side, the gap between the progress of AI development and institutional adaptation is the central problem. It is not a problem with a technical solution. It requires political will, international cooperation, and a willingness to accept that getting this right may require placing constraints on technological development that slow some of the benefits as well as some of the harms. The EU AI Act and the broader international governance effort represent notable progress, but they remain works in progress, and the forces pushing against effective regulation are substantial.

What I have tried to argue throughout this paper is that the questions about AI are not primarily technical. Rather, they are related with what kind of societies we want to be, what we value in human work and learning, how we want to govern the creation and distribution of knowledge, and what obligations more powerful institutions have to less powerful ones. AI makes those questions more urgent and more consequential. It does not answer them.

Disclosure: Generative artificial intelligence tools were used during the preparation of this manuscript to assist with language improvement. No AI system was used to generate scientific results, perform analyses, interpret data, or draw conclusions. All scientific content was developed, reviewed, and approved by the author, who take full responsibility of the manuscript.

References

Acemoglu, D., & Restrepo, P. (2018). *Artificial intelligence, automation, and work* (NBER Working Paper No. 24196). National Bureau of Economic Research.

Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and machine learning: Limitations and opportunities*. MIT Press.

Bengio, Y., et al. (2024). Managing extreme AI risks amid rapid progress. *Science, 384*(6698), 842–845.

Bommasani, R., et al. (2021). *On the opportunities and risks of foundation models*. arXiv preprint arXiv:2108.07258.

Brown, T., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Brynjolfsson, E., & McAfee, A. (2014). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. W. W. Norton.

Carr, N. (2010). *The shallows: What the Internet is doing to our brains*. W. W. Norton.

Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review, 107*(6), 1753–1820.

Clark, A. (2003). *Natural-born cyborgs: Minds, technologies, and the future of human intelligence*. Oxford University Press.

Dafoe, A. (2018). *AI governance: A research agenda*. Future of Humanity Institute, University of Oxford.

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv:1702.08608.

European Commission. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council (Artificial Intelligence Act).* Official Journal of the European Union.

Floridi, L., et al. (2018). AI4People: An ethical framework for a good AI society. *Minds and Machines, 28*(4), 689–707.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Harrington, E., Dhople, S., Wang, X., Choi, J., & Koester, S. (2022). Sustainability for semiconductors. *Issues in Science and Technology, 39*(1).

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence, 1*(9), 389–399.

Jumper, J., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583–589.

Kasneci, E., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274.

Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports, 15*, 17458.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems, 25*, 1097–1105.

Luckin, R. (2018). *Machine learning and human intelligence: The future of education for the 21st century*. UCL IOE Press.

McKinsey Global Institute. (2023). *The economic potential of generative AI: The next productivity frontier*. McKinsey & Company.

Obermeyer, Z., et al. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453.

OECD. (2023). *OECD employment outlook 2023: Artificial intelligence and the labour market*. OECD Publishing.

Pariser, E. (2011). *The filter bubble: What the Internet is hiding from you*. Penguin Press.

Patterson, D., et al. (2021). *Carbon emissions and large neural network training*. arXiv:2104.10350.

Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. *Nature Medicine, 25*(1), 37–43.

Rajpurkar, P., et al. (2022). AI in health and medicine. *Nature Medicine, 28*(1), 31–38.

Rolnick, D., et al. (2022). Tackling climate change with machine learning. *ACM Computing Surveys, 55*(2), 1–96.

Russell, S., & Norvig, P. (2020). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Scharre, P. (2018). *Army of none: Autonomous weapons and the future of war*. W. W. Norton.

Schwartz, R., et al. (2020). Green AI. *Communications of the ACM, 63*(12), 54–63.

Tison, G. H., et al. (2018). Passive detection of atrial fibrillation using a commercially available smartwatch. *JAMA Cardiology, 3*(5), 409–416.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460.

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society, 6*(1).

Wei, J., et al. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.

World Economic Forum. (2025). *The future of jobs report 2025*. WEF.

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

About the Author

Prof. Dr. Mehmet YILDIZ
Sabanci University | mehmet.yildiz[at]sabanci.edu.tr
0000-0003-1626-5858

Mehmet Yildiz is a Professor in the Materials Science and Nanoengineering and Manufacturing Engineering programs within the Faculty of Engineering and Natural Sciences at Sabanci University. Since 2018, he has been serving as the Vice President for Research. He earned his Ph.D. in Mechanical Engineering from the University of Victoria, Canada, in 2005, focusing on computational fluid dynamics and semiconductor single crystal growth. Following his Ph.D., he worked as a research associate and lecturer at the same university until 2007. He joined Sabanci University in 2007 as a faculty member in the Materials Science and Nanoengineering Program. Between 2013 and 2016, he established the Integrated Manufacturing Technologies Research and Application Centre (SUIMC) and its industrial stream, the Composite Technologies Centre of Excellence (CTCE), in collaboration with KORDSA. He served as the founding director of these centers until 2020, facilitating university-industry collaboration to accelerate technology maturation from TRL 1-3 to TRL 4-6. SU-IMC stands out as an open innovation center, uniquely integrating academic research with industrial applications, fostering interdisciplinary collaboration, and providing a dynamic ecosystem for startups, SMEs and OEMs to co-develop cutting-edge manufacturing solutions. He played a pivotal role in launching the Ph.D. program in Manufacturing Engineering in 2016. In addition to his academic responsibilities, he served as a deputy chairperson on KORDSA's board of directors for six years. He serves as a board member for Presidential Policy Board for Science, Technology, and Innovation of the Republic of Türkiye, and chairman of the Turkish Accelerator and Radiation Laboratory (TARLA). His research expertise spans advanced composite materials, nanocomposites, structural health monitoring, and computational mechanics and fluid dynamics, with a focus on meshless methods such as smoothed particle hydrodynamics and Peridynamics. He has authored over 260 indexed academic journal publications, 12 book chapters, and 178 conference presentations/papers. He has H-index of 53, and 8,080 citations according to Google Scholar. Additionally, he has filed over 10 patents and led over 70 national and international research projects, securing substantial research funding from national agencies, international programs, and industry partners. Prof. Yildiz has supervised more than 60 graduate students and mentored 33 postdoctoral fellows and visiting researchers across multiple disciplines, demonstrating a strong track record in both high-impact research and talent development.