SCIENTIFIC CULTURE

www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 2.1, (2026), pp. 13744-13751
Open Access. Online & Print

DOI: 10.5281/zenodo.20462576

HUMAN RESPONSIBILITY IN THE ERA OF ARTIFICIAL INTELLIGENCE (AI): CONTROL, ETHICS AND THE FUTURE

Huynh Minh Hau[1]

[1]University of Finance - Marketing, Ho Chi Minh City, Vietnam; Email: hmhau@ufm.edu.vn;
Orcid ID: https://orcid.org/0009-0006-4790-7648

Received: 04/04/2026
Accepted: 20/05/2026

Corresponding Author: Huynh Minh Hau
(hmhau@ufm.edu.vn)

ABSTRACT

The rapid advancement of Artificial Intelligence (AI) is fundamentally reshaping global economic, social, and cultural structures. Despite its immense benefits, AI introduces unprecedented challenges concerning ethics, safety, and control. This paper critically examines human responsibility in the AI era through three fundamental dimensions: (1) Technological control, (2) Ethical frameworks and (3) The responsibility of shaping the future. Employing a qualitative methodology and secondary data analysis of existing legal frameworks and the philosophy of technology, this study argues that AI is not an autonomous entity, but rather a reflection of its creators' values. Consequently, establishing a human-centric AI ecosystem requires proactive intervention from policymakers, engineers, and society. Ultimately, responsibility in the AI era cannot be delegated to machines; it relies entirely on human intellect and moral agency to ensure sustainable and symbiotic development.

KEYWORDS: Artificial Intelligence (AI), AI ethics, technological control, social responsibility, the future of humanity.

1. INTRODUCTION

We are entering a pivotal phase of the Fourth Industrial Revolution, wherein AI has transcended science fiction to become a ubiquitous catalyst across all facets of human life. From healthcare and education to finance and national security, AI delivers unprecedented operational efficiency. However, the proliferation of generative and autonomous AI models has precipitated profound existential inquiries: Who bears accountability when AI renders erroneous decisions? How can we ensure that AI remains aligned with human control? This article seeks to address these questions by elucidating the central role of human agency in AI development and deployment. In this context, responsibility transcends mere legal compliance; it necessitates a profound ethical consciousness and a strategic vision to navigate toward a sustainable human-machine symbiosis.

2. LITERATURE REVIEW

Over the past decade, and particularly following the proliferation of generative AI models, the discourse surrounding AI ethics and governance has garnered intense academic scrutiny. A comprehensive review of the literature reveals three primary schools of thought conceptualizing human responsibility toward AI: (1) The Value Alignment Problem, (2) Ethical principles for algorithmic design and (3) Human-machine interaction dynamics.

The Value Alignment Problem: Foundational research highlights the existential risk of misaligned objective functions in autonomous systems, warning against rigidly programmed goals given the fluidity of human values (Bostrom, 2014; Russell, 2019). Consequently, the focus has shifted toward bidirectional alignment. Recent scholarship in 2024 and 2025 argues that alignment must navigate complex moral trade-offs. For instance, Kneer and Viehoff (2025) conceptualize value forks in moral judgment, demonstrating empirically that human-AI alignment frequently confronts profound dilemmas between maximizing utility and ensuring fairness. Furthermore, Bradley and Saad (2024) underscore the intricate tension between aligning AI systems to human safety and the ethical treatment of advanced models, highlighting that contemporary AI alignment is a multidimensional, dynamic process rather than a static engineering task.

Algorithmic Ethics and the Fairness Accountability and Transparency (FAT) framework: The second research stream addresses immediate risks in narrow AI through the FAT framework. Machine learning algorithms often inherit and amplify systemic human biases in critical sectors (Park et al., 2021; Morris et al., 2024). With the formal enforcement of the European Union's AI Act in August 2024, the legal landscape for AI governance has fundamentally shifted. Contemporary legal and technological analyses emphasize that risk-based compliance, transparency mechanisms and fundamental rights protection must now be intrinsically embedded into algorithmic design from inception to avoid severe unintended consequences (Van Cleynenbreugel, 2025).

Human-in-the-Loop (HITL) and Moral Deskilling: To bridge the accountability gap inherent in black-box neural networks, the HITL model integrates human contextual judgment into the AI lifecycle. However, recent critical reviews argue that mere human oversight is insufficient. Salloch and Eriksen (2024) advocate for a transition from HITL to human-in-power, ensuring that humans remain the primary epistemic subjects rather than passive validators of machine outputs. If society continually delegates moral decisions to algorithms, it risks moral deskilling the erosion of humanity's intrinsic capacity for ethical reasoning (Vallor, 2015). Preserving human autonomy, therefore, requires proactively cultivating spaces for moral reflection rather than succumbing to cognitive automation.

While technical reports and philosophical essays provide valuable insights, they often remain fragmented. This study bridges the gap between algorithmic control imperatives and human ontology, demonstrating that the true risk lies in humanity's preparedness to shoulder the moral responsibility of shaping its technological future.

3. RESEARCH METHODOLOGY

In the context of the explosive advancement of AI, its ramifications are no longer confined to pure technical specifications but have permeated and profoundly impacted the philosophical, ethical, legal and operational structures of society. Consequently, the application of quantitative methods or cross-sectional empirical surveys would struggle to capture this systemic paradigm shift.

A qualitative research methodology, specifically through document analysis, enables the author to delve into the essence of technological phenomena, deconstruct complex streams of thought and comprehend regulatory discourses on a global scale. By inheriting, analyzing and systematizing preceding empirical works, strategic reports from international organizations and current legal frameworks, this study establishes a robust theoretical foundation to address existential

questions regarding control, ethical boundaries and human-machine symbiosis. This approach facilitates not only the identification of surface-level technological trends but more importantly the elucidation of why phenomena such as algorithmic black boxes or moral deskilling have emerged as pressing challenges to human responsibility in both the present and the future.

To ensure the comprehensiveness, objectivity and currency of a scientific article, the secondary data collection process was executed following a rigorous screening strategy based on three primary sources: Academic literature and philosophical monographs on technology: This category encompasses classic literature and peer-reviewed scientific articles indexed in prestigious international databases, focusing on AI safety theory, the value alignment problem, and digital humanism (e.g., the seminal works of Nick Bostrom, Stuart Russell, and Shannon Vallor). These materials provide profound ontological and technological-ethical depth.

International legal frameworks and technical standards: To accurately reflect the reality of technology governance, this study aggregates legal texts and policy proposals from pioneering legislative entities worldwide. The focal points include the European Union's AI Act and the ethical design guidelines promulgated by the Institute of Electrical and Electronics Engineers (IEEE). These empirical sources reflect humanity's concrete efforts in erecting technical and legislative guardrails for algorithms. Global strategic reports: The research integrates data from periodic reports by the World Economic Forum, particularly risk analyses concerning generative AI and the impact of automation on the global workforce. These documents supply macroeconomic and socio-demographic forecasting data, thereby bridging ethical theories with the dynamic realities of digital transformation.

The inclusion criteria mandate that sources be published by reputable organizations or scholars, possess verifiable scientific citations, and directly address the correlation between human control mechanisms and AI outcomes. Conversely, purely speculative literature lacking an empirical basis or commercial promotional materials from tech conglomerates were strictly excluded to safeguard the study's objectivity and high scientific rigor.

Following collection and categorization, the raw secondary data was processed, coded and rigorously analyzed utilizing three core logical techniques: synthesis, comparison and logical deduction. The Synthesis Method: This technique is applied to bridge the epistemological gap currently fragmenting two distinct disciplines: technical computer science and social philosophy. The author integrates technical concepts such as deep neural networks, objective functions and autonomous AI with moral constructs, including accountability, transparency, empathy and autonomy. This synthetic process shapes a comprehensive analytical framework for Human centric Responsibility, demonstrating that technical control and ethical orientation are two sides of the same coin in technology governance.

The Comparative Method: The study conducts a comparative analysis of divergent AI governance paradigms globally. Specifically, it contrasts the European Union's rigid, risk-based legislative approach with the voluntary, recommendatory ethical principles advocated by professional bodies like IEEE. This juxtaposition aims to delineate the merits and vulnerabilities of each paradigm, thereby deriving empirical lessons on constructing a legal corridor that concurrently guarantees ethical safety and fosters an ecosystem of innovation.

The Logical Deduction Method: This serves as the pivotal cognitive tool for linking contemporary realities to future scenarios. Proceeding from current empirical premises such as the black-box nature of algorithms rendering machine decision-making processes inexplicable to humans the study logically deduces the inevitable social corollary: should humanity continue to unconditionally delegate decision-making authority in high-risk domains (e.g., healthcare, jurisprudence, military) to machines, society will descend into a state of dissolved legal accountability and eroded intrinsic moral capacity. This method establishes a definitive causal nexus between present-day algorithmic design practices and the future trajectory of human existence.

To guarantee reliability and construct validity for this qualitative study, the author employs the data triangulation technique. By cross-referencing and corroborating a specific ethical argument or technical risk across all three independent sources (academic theory, normative legal texts and empirical economic reports), the study ensures its conclusions are objective, multidimensional and insulated from subjective authorial bias.

Nevertheless, the study acknowledges certain intrinsic methodological limitations. Given that AI is a technological domain characterized by extremely rapid turnover and breakthrough velocity, formally published academic literature occasionally exhibits a temporal lag relative to actual market developments. Furthermore, the analysis of secondary data is

contingent upon the availability and public disclosure of information by major tech conglomerates entities that currently monopolize the largest AI models. To mitigate these constraints, this study proactively and continuously incorporates the latest scientific preprints and technical white papers from leading research institutes, ensuring the article's scientific substance remains highly topical and possesses substantial applied value for journal readership.

4. DISCUSSION AND RESULTS

4.1. Controlling AI: Overcoming the black-box Barrier and the Accountability Conundrum in Critical Sectors

In the endeavor to mitigate technological risks, the scientific community currently confronts a colossal technical barrier: the black-box nature of neural networks and deep learning algorithms. Unlike traditional linear software systems, the analytical mechanisms of Artificial Intelligence, encompassing billions of parameters, are so profoundly complex that even their architects and programmers are occasionally powerless to explicitly articulate the rationale behind a specific output or decision. This profound deficit in interpretability engenders catastrophic risks when AI is deployed in the real world, particularly in high-stakes domains directly concerning human life and liberty.

Consequently, the imperative of accountability emerges as an immutable principle, necessitating that humans perpetually retain the role of the ultimate decision-maker in high-risk systems such as healthcare, jurisprudence and the military.

Practical paradigm in healthcare: The integration of Artificial Intelligence in diagnostic imaging and pathological analysis is yielding monumental strides. However, as analyzed by Topol (2019) regarding algorithmic medicine, should an erroneous AI diagnosis result in a fatal treatment protocol, the boundaries of legal liability blur precariously among the attending physician, the medical institution, and the software developer. If AI operates as an inexplicable black box, physicians cannot blindly adhere to its outputs. To neutralize this risk, the HITL model must be rigorously enforced. Within this framework, Artificial Intelligence serves strictly to conduct preliminary data analysis and propose scenarios, whereas the clinical physician synthesizes the context, renders the final verdict and bears ultimate legal responsibility.

Practical paradigm in the military and security: The advent of lethal autonomous weapons systems has incited profound global apprehension. The consensus among international organizations underscores that delegating the authority to terminate human life to an unfeeling algorithm constitutes a severe violation of international humanitarian law (Scharre, 2018). From a technical standpoint, engineers bear an immense responsibility to hardcode stringent safety constraints during the foundational training phase to preemptively neutralize any potential for the system to execute unforeseen, harmful autonomous actions.

4.2. AI Ethics: The Frontier of Fairness and the Amplification of Social Biases

The intrinsic nature of machine learning models lies in their capacity to learn and extract patterns from colossal data repositories historically generated by humans. Historical data inherently encapsulates a multitude of systemic injustices; consequently, AI inadvertently functions as a mirror, absorbing and even amplifying pre-existing societal prejudices pertaining to race, gender and religion. This necessitates a fundamental realization: the ethics of technology do not reside within the lines of programming code itself, but are entirely contingent upon the consciousness, vision and impartiality of the individuals executing the curation and classification of input data.

Practical paradigm of gender bias in recruitment: One of the most stark illustrations of this risk is the internal algorithmic recruitment failure at Amazon. According to Dastin (2018), Amazon endeavored to develop an AI system to automate the scoring and screening of candidate resumes. However, because the model was trained on the company's recruitment data over the preceding decade-an era predominantly dominated by men in the tech industry the system autonomously inferred that male candidates were inherently superior. Consequently, the AI systematically penalized or discarded resumes containing keywords associated with women or women's colleges. Amazon was ultimately forced to abandon the project upon realizing the impossibility of entirely eradicating the deeply entrenched bias within the model.

Practical paradigm of racial bias in the judicial system: Within the United States criminal justice system, the COMPAS algorithmic system has been widely utilized to assist judges in predicting the recidivism risk of defendants. Nevertheless, a landmark study by ProPublica (Larson et al., 2016) unmasked the system's severe algorithmic bias. The analytical findings demonstrated that COMPAS disproportionately and falsely assigned higher

recidivism risk scores to African American defendants, while frequently underestimating the risk for white defendants with comparable criminal histories. Reliance on a biased algorithm directly stripped numerous individuals of the opportunity for fair leniency, perpetuating a vicious cycle of systemic injustice.

To empirically validate the escalating risks of algorithmic bias, Table 1 delineates the exponential growth of AI-related ethical incidents and controversies, drawing on data from the Stanford AI Index Report.

Table 1. Growth of AI algorithmic abuse and bias incidents (2012–2023).

Evaluation Criteria	2012	2017	2021	2023
Cumulative AI ethical incidents recorded	12	108	250	> 500
Proportion of incidents involving discrimination and bias (%)	N/A	22%	45%	61%
Incidents in critical sectors (Healthcare, Justice, Security)	Low	Medium	High	Very High

Source: Synthesized and adapted by the author from the AI Index Report, Stanford University (Maslej et al., 2024).

This empirical trajectory underscores that as AI integration deepens in high-stakes domains, the amplification of systemic discrimination is not merely theoretical but a compounding real-world crisis, necessitating immediate human intervention in data curation.

From these evident repercussions, it can be affirmed that humanity cannot relegate social equity to machinery. Policymakers and civil society possess an urgent responsibility to establish robust legal corridors that safeguard privacy and ensure the transparency, diversity and objectivity of input datasets prior to their deployment in model training.

4.3. The Future: Comprehensive Preparation for a Symbiotic Generation and Labor Restructuring with AI

Diverging from extreme apocalyptic visions of machines entirely subjugating humanity, the practical trajectory of technological development indicates that humanity is laying the groundwork for a generation characterized by profound symbiosis between natural and artificial intelligence. The impact of AI on the global labor market is not purely destructive but structurally transformative. Specifically, artificial intelligence systems will not eradicate the future or entirely usurp human employment opportunities. Instead, a new mechanism of elimination predicated on competency will emerge individuals who master and adeptly integrate AI into their workflows will systematically replace the demographic that rejects or lacks the proficiency to interact with technology.

Practical paradigm of labor productivity shifts: In macroeconomic risk analyses, scholars have highlighted the dual-use nature and the dilemma of generative AI models wherein the technology simultaneously poses a risk of job displacement and acts as a colossal productivity lever if utilized optimally. A large-scale empirical study by Noy and Zhang (2023) at the Massachusetts Institute of Technology verified the impact of generative AI on knowledge workers. When assigned professional tasks such as data analysis and document drafting, the cohort utilizing AI assistance completed their workload in significantly less time. Concurrently, the quality of their output was independently rated as markedly superior to that of the cohort employing traditional manual methods. This empirical evidence bolsters the premise that AI tools function as cognitive augmentation rather than a complete substitution for the core value of human labor.

To safely and sustainably adapt to this novel economic architecture, educational infrastructure must undergo a systemic overhaul. Pedagogical models fixated on rote learning or repetitive skills have become obsolete. The educational system must rapidly pivot its focus toward cultivating critical thinking, fostering emotional intelligence and establishing a robust foundation of digital ethics for learners.

To substantiate the paradigm shift in labor and education, Table 2 delineates the projected skill demand trajectory forecasted by the World Economic Forum (2023). As illustrated, while basic cognitive tasks face aggressive automation, the demand for critical thinking and emotional intelligence is surging, reinforcing the premise that human-centric education remains the ultimate bulwark in the AI era.

Table 2. Projected shifts in core workforce skills under AI automation (2023–2027).

Skill CategoryNature of AI ImpactTraining Demand Trend (2023–2027)	Skill CategoryNature of AI ImpactTraining Demand Trend (2023–2027)	Skill CategoryNature of AI ImpactTraining Demand Trend (2023–2027)
Data Analysis & Computational Thinking	Strongly augmented by AI	High growth (+30%)

Analytical & Creative Critical Thinking	Irreplaceable	Top educational priority (+73%)
Reading, Basic Math & Manual Recording	Highly susceptible to automation	Significant decline (-25%)
Emotional Intelligence & Ethics	Human autonomy and moral agency	Mandatory for high-risk sectors

Source: Systematized by the author based on the Future of Jobs Report, World Economic Forum (WEF, 2023).

As illustrated, while routine cognitive tasks face aggressive automation, the demand for critical thinking, emotional intelligence, and ethical judgment is surging. This reinforces the premise that cultivating humanity's moral muscle remains the ultimate educational bulwark in forging a sustainable symbiotic future.

Practical paradigm of national educational strategy: A pioneering nation in universalizing AI symbiotic skills is Finland. Rather than confining computer science knowledge to the narrow purview of technical universities, the Finnish government implemented a national educational strategy designed to equip the entire populace with foundational knowledge of algorithms and machine learning mechanisms (Vincent-Lancrin & van der Vlies, 2020). This strategy exemplifies proactive state intervention, enabling the public to dismantle psychological barriers of technophobia while acquiring sufficient critical capacity to confront and process automated decisions in daily life and production.

Alongside efforts from the educational sector, macroeconomic intervention from regulatory bodies and economic institutions is a prerequisite for ensuring a humanistic digital transformation. Government agencies and the corporate sector must transparently exhibit their social responsibility by establishing financial support funds and deploying targeted reskilling and upskilling programs for vulnerable labor demographics directly impacted by the wave of automation.

Practical paradigm of corporate social responsibility: Within the private economic sector, Amazon's commitment to reskilling its workforce serves as a quintessential paradigm of executing social responsibility to resolve the labor redundancy conundrum (Bughin et al., 2018). Rather than executing mass layoffs of warehouse workers upon the integration of robotics and automated sorting systems, the corporation allocated a massive budget to retrain hundreds of thousands of employees. Through this initiative, unskilled laborers were supported in transitioning to higher value-added roles, such as system maintenance technicians or data flow management specialists. This maneuver not only mitigates the unemployment shock induced by technology but also manifests the organization's paramount ethical responsibility toward the stability of social welfare.

5. CONCLUSION

Artificial Intelligence (AI) represents one of the most monumental leaps in the history of human technology, fundamentally reshaping economic, social, and cultural structures on a global scale. Through a comprehensive analysis integrating algorithmic engineering and social philosophy, this study affirms that the developmental trajectory of AI is neither a deterministic process nor an independent destiny. The core essence of artificial intelligence remains a reflection, retention, and amplification of the values, worldviews, and biases inherent to its human creators. Consequently, architecting a human-centric Artificial Intelligence ecosystem demands a collective awakening and proactive intervention from all societal sectors. Ultimately, technological risks pertaining to safety, ethics, and control are the direct corollaries of contemporary human choices and responsibilities; they cannot be offloaded onto or blamed on machines. The endeavor to forge a humanistic future, characterized by a sustainable human-machine symbiosis, relies entirely upon our resolve to confront and resolve the foundational challenges of our era.

This study has elucidated the technical and systemic challenges humanity faces in maintaining control over generative and autonomous AI models. The black-box nature of deep learning algorithms and complex neural networks erects a formidable barrier, rendering even programmers occasionally incapable of interpreting or reverse-engineering computational decision-making processes. This interpretability deficit dictates that humans must unequivocally retain the ultimate decision-making authority in high-stakes domains such as healthcare, jurisprudence, and the military through the rigorous implementation of the HITL model. Systemic value alignment and the hardcoding of safety constraints during the initial training phase are mandatory prerequisites for preventing harmful autonomous behaviors (Bostrom, 2014; Russell, 2019). Absent this cognitive oversight, blind delegation to algorithms will precipitate existential catastrophes wherein the machine's optimization goals radically diverge from humanity's core norms and interests.

Parallel to the imperative of technical control is the intricate frontier of algorithmic ethics and social

equity. Trained on colossal datasets that mirror historical societal structures, AI is acutely susceptible to absorbing and amplifying systemic biases and discrimination regarding gender, race and religion. Empirical evidence from human resource recruitment and judicial scoring systems demonstrates that bias does not originate from mindless lines of code, but from the consciousness of those curating the input data. It is a human responsibility to institute robust regulatory frameworks and transparency standards anchored in FAT principles to safeguard privacy and equity (European Commission, 2021). Furthermore, on a profound philosophical level, over-reliance on automated AI judgments engenders the risk of human moral deskilling (Vallor, 2015). If moral capacity, empathy, and critical thinking are not consistently exercised through real-world dilemmas, humanity's moral muscle will atrophy, ultimately stripping the species of its autonomy and moral agency.

Looking toward the horizon, the AI revolution does not eradicate human opportunity but rather restructures the entire labor market and cognitive landscape. The paradigm of productivity displacement indicates that AI will not absolutely usurp human employment; however, labor cohorts adept at working symbiotically with technology will swiftly supersede those lacking interactive proficiency (Bughin et al., 2018; Noy & Zhang, 2023). To prepare for this symbiotic generation, the educational system must pivot from traditional knowledge transmission toward the cultivation of emotional intelligence, critical thinking, and digital ethics. Concurrently, the automation transition necessitates meticulous societal preparation via reskilling programs designed to protect vulnerable demographics, thereby transmuting AI from a menacing tool into an augmentative force that elevates human productivity and humanistic value.

Although this study constructs an integrated and comprehensive analytical framework, this qualitative research predicated on secondary document analysis-acknowledges intrinsic limitations due to the temporal lag of academic publications relative to the velocity of real-world technological breakthroughs. The prospective emergence of Artificial General Intelligence will inevitably introduce novel risk scenarios transcending the scope of current legal frameworks. Therefore, future research trajectories should focus on the empirical measurement of algorithmic bias across specific cultural datasets, while simultaneously conducting extensive empirical surveys on the cognitive impact of AI on the moral decision-making capacities of professionals in critical sectors. Only by continuously updating our knowledge and courageously shouldering the responsibility to shape technology can humanity preserve its core humanistic values in the era of artificial intelligence.

AUTHOR CONTRIBUTIONS: Conceptualization, H.M.H.; methodology, H.M.H.; formal analysis, H.M.H.; investigation, H.M.H.; resources, H.M.H.; data curation, H.M.H.; writing - original draft preparation, H.M.H.; writing - review and editing, H.M.H. The author has read and agreed to the published version of the manuscript.

ACKNOWLEDGMENTS: This research is funded by the University of Finance - Marketing, Vietnam. The author, Huynh Minh Hau, would like to express sincere gratitude for the financial support provided by the University. The author also thanks the anonymous reviewers and the editorial board for their constructive comments and suggestions, which helped improve the quality of this paper.

REFERENCES

Bostrom, N. (2014) Superintelligence: Paths, Dangers, Strategies. Oxford, Oxford University Press.

Bradley, A., & Saad, B. (2024). AI alignment vs AI ethical treatment: Ten challenges. Global Priorities Institute Working Paper Series, No. 19-2024. University of Oxford.

Bughin, J., Hazan, E., Lund, S., Dahlström, P., Wiesinger, A., & Subramaniam, A. (2018). Skill shift: Automation and the future of the workforce. McKinsey Global Institute. https://www.mckinsey.com/featured-insights/future-of-work/skill-shift-automation-and-the-future-of-the-workforce

Dastin, J. (2018, October 11). Amazon scraps secret AI recruiting tool that showed bias against women. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

European Commission (2021) Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Brussels, European Commission.

Gabriel, I. (2020). Artificial intelligence, values, and alignment. Minds and Machines, 30(3), 411-437. https://doi.org/10.1007/s11023-020-09539-2

Goyal, N., Chang, Y.-C., Li, Y., Lin, Y.-C., Huang, Y., & Zhang, Y. (2024). Towards bidirectional human-AI alignment: A systematic review for clarifications, framework, and future directions. arXiv preprint arXiv:2406.09264.

Hendrycks, D., Carlini, N., Schulman, J., & Steinhardt, J. (2021). Unsolved problems in ML safety. arXiv preprint arXiv:2109.13916.

IEEE (2019) Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems. IEEE Global Initiative.

IEEE Global Initiative. (2019). Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems. IEEE.

Kneer, M., & Viehoff, J. (2025). The Hard Problem of AI Alignment: Value Forks in Moral Judgment. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT 2025) (pp. 2671-2681). Association for Computing Machinery. https://doi.org/10.1145/3715275.3732174

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. ProPublica. https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

Maslej, N., Fattorini, L., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Ngo, H., Niebles, J. C., Parli, V., Shoham, Y., Wald, R., Clark, J., & Perrault, R. (2024). The AI index 2024 annual report. Stanford University: AI Index Steering Committee, Institute for Human-Centered AI.

Morris, M. R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., ... & Legg, S. (2024). Levels of AGI: Operationalizing progress on the path to artificial general intelligence. arXiv preprint arXiv:2311.02462.

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. Science, 381(6654), 187-192. https://doi.org/10.1126/science.adh2586

Park, J., Lee, J., & Kim, H. (2021). Algorithmic bias in hiring AI: A review and framework for mitigation. Journal of Business Ethics, 1-18.

Russell, S. (2019) Human Compatible: Artificial Intelligence and the Problem of Control. New York, Viking.

Salloch, S., & Eriksen, T. E. (2024). From Human-in-the-loop to Human-in-power: Emphasizing the Epistemic Agency of Patients in Artificial Intelligence-assisted Medicine. Bioethics, 38(7), 1-9.

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

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44-56. https://doi.org/10.1038/s41591-018-0300-7

Vallor, S. (2015). Moral deskilling and upskilling in a new machine age: Reflections on the ambiguous future of character. Philosophy & Technology, 28(1), 107-124. https://doi.org/10.1007/s13347-014-0156-9

Van Cleynenbreugel, P. (2025). The EU AI Act: Law of Unintended Consequences?. Technology and Regulation (TechReg), 2025(016), 316-335.

Vincent-Lancrin, S., & van der Vlies, R. (2020). Trustworthy artificial intelligence (AI) in education: Promises and challenges. OECD Education Working Papers, No. 218. OECD Publishing. https://doi.org/10.1787/a6c90fa9-en

World Economic Forum (WEF) (2023) Global Risks Report: The dual-use dilemma of generative AI. Geneva, Switzerland.

World Economic Forum (WEF). (2023). The future of jobs report 2023. Geneva, Switzerland. https://www.weforum.org/reports/the-future-of-jobs-report-2023