RESPONSIBLE AI UNDER THE PHILOSOPHICAL FRAMEWORK OF DIGITAL HUMANISM

Fei Sun, Damir Isovic

***Abstract**: As Artificial Intelligence (AI) becomes increasingly influential, ethical concerns surrounding its impact on society and people grow. Current AI ethics frameworks often prioritize efficiency and optimization, neglecting core humanistic values such as dignity, autonomy, and social justice. Digital Humanism offers an essential philosophical foundation to ensure that AI aligns with human values and societal well-being. This paper explores the integration of Responsible AI within the philosophical framework of Digital humanism. It argues that Digital Humanism principles, such as Human-Centered Design, inclusivity, transparency, and ethical universalism, offer critical guidelines for developing and deploying AI that serves humanity. In doing so, this work advocates a humanistic approach to AI development that transcends the limitations of technocentric paradigms.*

***Keywords**: Responsible AI, Digital Humanism, Ethics AI, Human-Centered Design, Contextual Understanding in AI.*

***ITHEA Keywords**: Computers and Society, Philosophy of Computing, Computer Ethics*

***DOI**: https://doi.org/10.54521/ijita32-04-p04*

Introduction

AI has become an integral part of our lives, shaping how we communicate, travel, work, shop, learn, and access services—deeply embedded across nearly every aspect of modern society. While AI promises numerous benefits, its rapid advancement raises critical concerns regarding fairness, accountability,

transparency, and its broader societal impact. This growing demand for Responsible AI [RAI]—defined as the development and deployment of AI that integrates ethical, legal, and social considerations—has never been more pressing [Dignum, 2019].

Current AI ethics frameworks, which emphasize principles like fairness, transparency, and explainability, are often grounded in a technocentric paradigm. These frameworks prioritize efficiency and optimization over core human values. Digital Humanism (DigHum) offers a necessary perspective shift, focusing on human dignity, autonomy, and cultural diversity as central to AI governance [Werthner et al., 2021]. This paper argues that DigHum provides a solid and comprehensive philosophical foundation for RAI. By grounding AI development in humanistic principles, DigHum challenges reductionist models that reduce humans to mere data points or algorithmic entities [Schultz et al., 2024], ensuring that AI serves human interests and promotes societal well-being.

Philosophical Principles of Digital Humanism

DigHum seeks to shape technology in ways that uphold human dignity, autonomy, and societal values. Rooted in Enlightenment ideals—i.e., rationality, autonomy, and individual dignity [Nida-Rümelin & Winter, 2023]—it rejects two key stances: one that reduces humans to mere machines [mechanistic] and another that attributes human qualities to machines [animistic] [Nida-Rümelin & Staudacher, 2023]. Instead, DigHum asserts that technology should enhance human life and align with core human values, rather than undermine or replace them. This perspective emphasizes that technology should contribute to human development and well-being.

By amalgamating these core humanistic principles into the development and application of digital technologies, DigHum ensures that technology serves humanity's best interests. Its philosophical foundation is explored in the key texts referenced above.

Human-Centered Design and Cultural Sensitivity

A core tenet of Digital Humanism is that technological innovations must prioritize human well-being and dignity, while respecting the diversity of cultures, contexts, and individual needs. Digital solutions should not be designed to impose a universal approach but be adaptable to different societal, cultural, and personal circumstances. Digital creations must reflect the richness of human diversity, avoiding the sustainment and reinforcement of harmful stereotypes or biases. Similarly, Nida-Rümelin and Winter's Humanism and Enlightenment advocates technologies that acknowledge the varying experiences of individuals across cultures, and foster inclusivity and equity. Therefore, tech designers need to be more aware of their own biases to ensure that their creations are inclusive and equitable.

Democracy, Inclusion, and Ethical Universalism

Technologies should be developed to combat inequality and to ensure that their benefits are accessible to all, irrespective of social or economic status. The Vienna Manifesto on Digital Humanism [2019] underscores that technological progress should align with human rights and democratic principles. Humanism and Enlightenment also stresses that digital innovations should be rooted in ethical universalism, ensuring that they benefit all people, promote justice, and respect individual dignity, while addressing societal challenges in a fair and inclusive manner.

Rational Inquiry and Thoughtful Development

At the heart of Digital Humanism lies the Enlightenment principle of rational inquiry, which calls for critical reasoning and scientific reflection in the development of digital innovations. Both Humanism and Enlightenment and Philosophical Foundations of Digital Humanism [Nida-Rümelin & Staudacher, 2023] emphasize the importance of thoughtful, inquiry-driven development. By applying critical analysis and rational deliberation, digital creations can be designed to align with human interests, avoid harm, and adhere to ethical guidelines. Rationality ensures that these technologies serve the public good while maintaining accountability to society.

Transparency, Accountability, and Ethical Oversight

In accordance with Enlightenment ideals, Digital Humanism advocates transparency and accountability in the design and operation of digital systems. The Vienna Manifesto on Digital Humanism calls for clear regulatory frameworks to ensure the openness of technological processes and hold developers accountable for the impact of their innovations. Humanism and Enlightenment further emphasizes the need for these tools to be transparent in their functions, allowing for external scrutiny and fostering public trust. Ethical oversight ensures that digital creations are aligned with societal values and contribute to the common good without undermining individual rights or freedoms.

Human Leadership and Social Responsibility

A core principle of Digital Humanism is maintaining human authorship and responsibility in technological progress. Nida-Rümelin and Staudacher [2023] stress that humans must control the tools they create, ensuring technology serves societal goals. Nida-Rümelin and Winter [2023] highlight the social responsibility of technologists, urging them to prioritize ethics and broader impacts. This principle upholds human values in digital systems, ensuring accountability. As stated in Vienna Manifesto on Digital Humanism, decision-makers must remain responsible, with automated systems supporting—not replacing—human decision-making.

Regulation, Education, and Ethical Frameworks

Finally, Digital Humanism advocates for the establishment of regulatory frameworks that guide the ethical development of digital tools and systems. The Vienna Manifesto stresses the importance of creating structures that promote fairness, accountability, and transparency. Additionally, Philosophical Foundations of Digital Humanism by Nida-Rümelin and Staudacher calls for the integration of ethics into educational curricula, especially in the STEM fields, to prepare future technologists to understand the societal implications of their work. By incorporating ethics into education and policy, Digital Humanism future

technological innovators will be equipped to build systems that align with societal needs and moral principles.

Why Digital Humanism Matters for AI

Recent advances in generative AI have significantly raised public awareness of the relation between technology and society, as well as concerns regarding its potential threats. Human and Societally Critical [HSC] systems increasingly include autonomous agents, machine-learning trained components, or even AI-generated software from natural language descriptions. These substantial changes present additional challenges for software development, especially in HSC scenarios. Furthermore, AI enables the development of larger and more complex systems that may exhibit unanticipated behavior. Complete control over such systems is unattainable, and the illusion of control that software developers usually have can be even more dangerous [Lee, 2024].

Neglecting the human and societal dimensions of advanced technological developments can result in significant consequences, particularly due to the rapid pace of innovation. Ideas can be swiftly converted into disruptive products, often without adequate time for thorough validation.

Digital Humanism is crucial for AI to address the challenges above, as well as current development that prioritizes profit, emphasizing efficiency over human well-being [Preston, 2022]. This technocentric focus favors fairness and explainability metrics while overlooking broader societal impacts like democracy, autonomy, and social justice. Consequently, AI systems risk becoming reductionist, oversimplifying complex human behaviors [Yampolskiy, 2025], and perpetuating algorithmic biases by failing to account for evolving cultural and social contexts. Additionally, accountability remains fragmented among developers, regulators, and users, creating ambiguity in ethical responsibility for AI's societal consequences.

Digital Humanism as Philosophical Foundation in Responsible AI

Digital Humanism offers a solid philosophical foundation for Responsible AI by embedding human-centered values into AI development. This approach ensures that technology aligns with human dignity, autonomy, and societal well-being. It provides a structured way to design AI systems that are ethical, inclusive, and accountable. A key contribution of Digital Humanism to RAI focuses on human-centered AI design. Instead of replacing or controlling human life, AI should enhance it, creating systems that respect cultural diversity, individual rights, and social contexts, and avoid biases that marginalize certain groups. E.g. AI in healthcare and education needs to adapt to different cultural, ethical and ideological perspectives, rather than to enforce a one-size-fits-all model.

Digital Humanism also reinforces fairness and justice in AI governance. It advocates democratic participation and ethical universalism, ensuring AI development is inclusive and benefits society rather than deepening inequalities. AI systems should reduce biases, promote equitable access, and prioritize social good over pure economic incentives.

Interdisciplinary collaboration is crucial for RAI, as AI development should integrate expertise from computer science, ethics, sociology, law, and the humanities to ensure AI is both technically robust and socially responsible. However, interdisciplinary work can be challenging due to factors such as differing objectives, varying terminologies, and methodological incompatibilities. Addressing these issues systematically requires a shared framework of values and priorities, as proposed by DigHum. DigHum prevents AI ethics from being reduced to mere mathematical optimization or regulatory compliance, instead highlighting broader human and societal impacts.

Transparency and accountability are essential to align AI with human values. Digital Humanism calls for AI systems to be understandable, explainable, and subject to oversight, preventing them from becoming opaque "black boxes" operating beyond human control.

Embedding transparency and accountability mechanisms ensures that AI decision-making adheres to ethical and democratic principles. Digital Humanism posits that AI must remain under human control and serve human interests. It should function as an augmentation of human intelligence, guided by ethical and social responsibilities and human values, rather than as an autonomous decision-maker. Maintaining human oversight safeguards against unintended consequences and prevents AI from being misused for manipulation, surveillance, or control without ethical justification.

Digital Humanism also highlights the role of regulation and ethical education in shaping responsible AI. Policymaking must uphold humanistic values by enforcing transparency, fairness, and accountability in AI systems. Additionally, ethics should be integrated into STEM education to equip future developers with tools to design AI that aligns with human rights and democratic values.

By applying Digital Humanism as a framework, RAI can move beyond technical and regulatory solutions to embrace a deeper ethical foundation, ensuring AI that not only functions efficiently, but also actively contributes to a more just, inclusive, and human-centered society.

Challenges and Conclusion

Integrating Digital Humanism into Responsible AI development entails several challenges. Encoding abstract ethical values like fairness and justice into algorithms is difficult, as these are context-dependent and culturally variable [Pery, Rafiei, Simon, & van der Aalst, 2022] [Hung & Yen, 2023]. Besides, institutional barriers, prioritizing profit over ethical considerations [De Cremer & Narayanan, 2023], hinder the adoption of a human-centered approach to AI. AI technology's global nature also poses challenges in establishing universal ethical standards.

Despite these challenges, Digital Humanism principles offer a promising pathway toward a more responsible, human-centered AI. Through interdisciplinary collaboration, participatory design, and commitment to ethical pluralism, AI can be developed in a way that reflects our highest human

aspirations. By grounding AI in humanistic values, we can ensure that it serves the public good, fosters human flourishing, and contributes to a more just and equitable society.

Bibliography

[De Cremer and Narayanan, 2023] De Cremer, D., Narayanan, D. On educating ethics in the AI era: why business schools need to move beyond digital upskilling, towards ethical upskilling. AI and Ethics, Vol.3, Springer, 2023. pp. 1037–1041 https://doi.org/10.1007/s43681-023-00306-4

[Dignum, 2019] Dignum, V. Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way. Springer Nature, 2019. [Hung and Yen, 2023] Hung, T.-W., Yen, C.-P. Predictive policing and algorithmic fairness. Synthese, Vol.201, Issue 206. Springer, 2023. https://doi.org/10.1007/s11229-023-04189-0

[Hung & Yen, 2023] Hung, T.-W., & Yen, C.-P. Predictive policing and algorithmic fairness. Synthese, Vol. 201, No. 6, Springer, 2023. Article 206. https://doi.org/10.1007/s11229-023-04189-0

[Lee, 2024] Lee, E.A. Are We in Control? In: Springer Nature Switzerland, Cham, 2024. pp. 165–174. https://doi.org/10.1007/978-3-031-45304-5_11

[Nida-Rümelin and Staudacher, 2023] Nida-Rümelin, J., Staudacher, K. Philosophical Foundations of Digital Humanism. In: Werthner, H. et al. [Eds.]. Introduction to Digital Humanism. Springer, 2023. pp. 17–30. https://doi.org/10.1007/978-3-031-45304-5_2

[Nida-Rümelin and Winter, 2023] Nida-Rümelin, J., Winter, D. Humanism and Enlightenment. In: Werthner, H. et al. [Eds.]. Introduction to Digital Humanism. Springer, 2023. pp. 3–16.https://doi.org/10.1007/978-3-031-45304-5_1

[Pery et al, 2022] Pery, A., Rafiei, M., Simon, M., van der Aalst, W.M.P. Trustworthy Artificial Intelligence and Process Mining: Challenges and

Opportunities. In: Process Mining Workshops, ICPM 2021. Springer, Cham, 2022. pp. 395–407. https://doi.org/10.1007/978-3-030-98773-1_30

[Preston, 2022] Preston, J. Artificial intelligence in the capitalist university: Academic labour, commodification, and value. Routledge, 2022. https://doi.org/10.4324/9781003081654

[Schultz et al, 2024] Schultz, M.D., Clegg, M., Hofstetter, R., et al. Algorithms and dehumanization: a definition and avoidance model. AI & Society, Springer,2024. https://doi.org/10.1007/s00146-024-02123-7

[Vienna Manifesto, 2019] Vienna Manifesto on Digital Humanism. Vienna Manifesto on Digital Humanism. 2019. Retrieved from Digital_Humanism_Manifesto-June_2019-FINAL.indd

[Yampolskiy, 2025] Yampolskiy, R.V. On monitorability of AI. AI and Ethics, Vol.5, Springer, 2025. pp. 689–707. https://doi.org/10.1007/s43681-024-00420-x

Author's Information

Fei Sun – *Mälardalen University, School of Innovation, Design and Engineering; Ph.D student in Computer Science. Box 883, 721 23 Västerås Sweden;*

e-mail: fei.sun@mdu.se

Major Fields of Scientific Research: Responsible AI, Digital Humanism

Damir Isovic *– Mälardalen University, School of Innovation, Design and Engineering; Associated Professor in Computer Science. Box 883, 721 23 Västerås Sweden;*

e-mail: damir.isovic@mdu.se

Major Fields of Scientific Research: Real-Time Systems, Digital Humanism, Quantum Computing