A Value-Based Approach to AI Ethics: Accountability, Transparency, Explainability, and Usability

Un Enfoque Basado en Valores para la Ética de la Inteligencia Artificial: Responsabilidad, Transparencia, Explicabilidad y Usabilidad

Vish Iyer
University of Northern Colorado, Estados Unidos de América
vish.iyer@unco.edu
Moe Manshad
University of Northern Colorado, Estados Unidos de América
muhanad.manshad@unco.edu
https://orcid.org/0000-0003-4068-8850
Daniel Brannon
University of Northern Colorado, Estados Unidos de América
daniel.brannon@unco.edu
https://orcid.org/0000-0002-1100-6788

Recepción: 19 Agosto 2024
Aprobación: 12 Diciembre 2024

Acceso abierto diamante

Abstract

As artificial intelligence (AI) becomes increasingly prevalent, ensuring its ethical development and deployment is paramount. This paper proposes a value-based approach to AI ethics, focusing on four key principles: accountability, transparency, explainability, and usability. By examining these principles, providing real-world examples, and discussing implementation challenges, we contribute to the ongoing discourse on responsible AI development and offer practical insights for stakeholders across various industries.
JEL code: O30, K24, D63
Keywords: Artificial Intelligence Ethics, Accountability, Transparency, Explainability, Usability.

Resumen

A medida que la inteligencia artificial (IA) se vuelve cada vez más omnipresente en la sociedad, garantizar su desarrollo e implementación ética es fundamental. Este documento propone un enfoque basado en valores para la ética de la IA, centrándose en cuatro principios clave: responsabilidad, transparencia, explicabilidad y usabilidad. Al examinar estos principios a través de una revisión bibliográfica exhaustiva y proporcionar ejemplos del mundo real, contribuimos al discurso continuo sobre el desarrollo responsable de la IA y ofrecemos ideas prácticas para las partes interesadas de diversas industrias.
Código JEL: M21.
Palabras clave: Ética de la Inteligencia Artificial, Responsabilidad, Transparencia, Explicabilidad, Usabilidad.

INTRODUCTION

The rapid advancement and integration of artificial intelligence (AI) into various aspects of society have brought unprecedented opportunities and challenges (Bostrom, 2014). As AI systems increasingly influence decision-making processes in critical domains such as healthcare, finance, and governance, ensuring their ethical development and deployment has become crucial (Jobin et al., 2019). This paper proposes a value-based approach to AI ethics, focusing on four fundamental principles: accountability, transparency, explainability, and usability.

The potential impacts of AI on society are profound. As Bostrom and Yudkowsky note, advanced AI systems could have far-reaching consequences on human life, potentially reshaping economies, social structures, and humanity's future. Therefore, we must develop and deploy AI systems that align with human values and ethical principles (Bostrom & Yudkowsky).

THE VALUE-BASED APPROACH TO AI ETHICS

A value-based approach to AI ethics entails grounding AI systems' development, deployment, and use in core ethical values (Dignum, 2018). This approach aims to create AI systems that are not only technically proficient but also aligned with societal values and moral standards. By prioritizing accountability, transparency, explainability, and usability, we can foster responsible AI usage and mitigate potential risks associated with AI technologies.

The European Commission's High-Level Expert Group on Artificial Intelligence emphasizes the importance of this approach, stating that trustworthy AI should be lawful, ethical, and robust. They argue that ethical AI is crucial for ensuring that AI systems respect fundamental rights, societal values, and ethical principles (European Commission's High-Level Expert Group on Artificial Intelligence).

THEORETICAL FRAMEWORK: VALUE-BASED AI ETHICS

Conceptual Foundations

A value-based approach to AI ethics transcends traditional technological considerations, embedding ethical principles into the core of technological design. As Hagendorff (2020) critically evaluates, existing AI ethics guidelines often need to provide comprehensive ethical frameworks, necessitating a more nuanced approach to technological governance.

Ethical Landscape and Global Perspectives

Drawing from an extensive analysis of international research, we identify four fundamental ethical principles. Wong and Cheung's (2022) comparative study of global AI regulation underscores the importance of developing adaptable, context-sensitive ethical approaches to navigate the complex international technological landscape.

Accountability

Accountability in AI refers to the responsibility of individuals and organizations for the outcomes of AI systems. It involves establishing clear governance structures, conducting ethical impact assessments, and implementing continuous monitoring mechanisms (IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems). Accountability ensures that AI developers and deployers are answerable for their systems' decisions and actions, promoting trust and responsible innovation.

Example of Accountability in AI

Consider the case of an AI-powered hiring system used by a large corporation. To ensure accountability:

The company establishes a transparent chain of responsibility, designating specific teams and individuals responsible for the AI system's decisions. Regular audits are conducted to assess the system's performance and identify any biases in hiring decisions. The company implements a mechanism for candidates to contest decisions made by the AI system, ensuring human oversight and the ability to correct errors. The development team regularly reports to a diverse ethics board that includes external stakeholders, ensuring broader societal perspectives are considered.

This approach aligns with the recommendations of Gupta et al. (2018), who emphasize the importance of verifiable claims about AI systems' behavior and impact to build trust and accountability (Gupta et al., 2018).

Accountability in AI extends beyond traditional responsibility mechanisms. Selbst and Powles (2018) highlight the critical importance of developing meaningful information disclosure frameworks that enable genuine understanding and oversight.

Key accountability strategies include: Transparent decision-making processes, Comprehensive impact assessments, Mechanisms for Algorithmic Contestability

Transparency

Transparency in AI involves making AI systems' functionality, decision-making processes, and potential biases accessible and understandable to stakeholders (European Commission's High-Level Expert Group on Artificial Intelligence, 2019). This principle is crucial for building trust in AI technologies and enabling meaningful oversight. Transparency includes disclosing data sources, algorithmic processes, and potential societal impacts of AI systems.

Example of Transparency in AI

Let us consider a predictive policing AI system used by a city's police department: The police department publicly discloses the data sources used to train the AI, including historical crime data and demographic information. The department clearly explains how the AI system weighs different factors to predict potential crime hotspots.

Regular reports show the system's accuracy rates and any discrepancies in predictions across different neighborhoods or demographic groups. The algorithmic model is available for independent audits by academic researchers and civil rights organizations. This level of transparency allows for public scrutiny. It helps identify potential biases or unintended consequences, as emphasized by Floridi et al. in their ethical framework for a good AI society (Floridi et al., 2018).

Transparency emerges as a crucial mechanism for building public trust. Green's (2019) research on institutional accountability provides insights into how complex technological systems can develop trust through deliberate, comprehensive disclosure mechanisms Green (2019).

Explainability

Explainability pertains to the ability to elucidate and justify the rationale behind AI-generated decisions (Brundage et al., 2018). This principle is particularly critical in high-stakes domains where AI-informed choices can have significant consequences. Explainable AI systems allow users and affected parties to understand the basis of AI-generated outputs, facilitating informed decision-making and contestability.

Example of Explainability in AI

Consider an AI system used in healthcare for diagnosing diseases: The AI provides a diagnosis and highlights the specific symptoms, test results, and factors in the patient's history that led to its conclusion. The system uses visualization techniques to show which areas of medical images (e.g., X-rays or MRIs) were most influential in its diagnosis.

The AI provides a confidence score and alternative possibilities for each diagnosis, helping doctors understand the certainty of the AI's decision. The system can generate natural language explanations of its reasoning process, tailored to medical professionals and patients.

This approach to explainability aligns with the recommendations of Amodei et al., who highlight the importance of interpretable AI systems in ensuring safety and reliability (Amodei et al., 2016). The challenge

of explainability is particularly acute in high-stakes domains. De Vries (2020) examines the critical role of explainability in medical AI, demonstrating how transparent decision-making processes can mitigate potential risks and build professional trust.

Usability

Usability in AI encompasses ensuring that AI interfaces and outputs are user-friendly, intuitive, and effective in meeting the needs of their intended users (Fjeld et al. 2020). This principle is vital for promoting the practical application of AI insights and recommendations. Usable AI systems consider accessibility, inclusivity, and user empowerment, enabling diverse user groups to interact effectively with AI technologies.

Example of Usability in AI

Let us examine a personal finance AI assistant: The AI uses natural language processing to allow users to interact with it using everyday language rather than requiring specific commands. The interface is designed to be accessible to users with disabilities, including screen reader compatibility and voice control options.

The AI adapts its communication style and complexity based on the user's financial literacy level and preferences. The system provides straightforward, actionable suggestions for improving economic health, with step-by-step implementation guidance. Users can easily customize the AI's focus areas and the frequency and type of notifications they receive.

This focus on usability aligns with the "Designing AI for Social Good" principles outlined by Fjeld et al., emphasizing the importance of inclusivity and user empowerment in AI systems (Fjeld et al., 2020). Usability transcends traditional interface design. Van Dijck and Poell's (2021) research on social media platforms illustrates how AI-driven technologies transform contemporary societal interactions, emphasizing the need for inclusive, adaptable design principles.

Implementing the Value-Based Approach

Implementing a value-based approach to AI ethics requires concerted efforts from various stakeholders, including developers, policymakers, and end-users. Key strategies include:

Developing ethical guidelines and governance frameworks operationalizing these principles (Morley et al. 2020).

Incorporating ethical considerations into the AI development lifecycle, from design to deployment and monitoring (Gupta et al., 2018). Fostering interdisciplinary collaboration to address the complex ethical challenges AI technologies pose (Rahwan, 2018). Promoting education and awareness about AI ethics among developers, users, and the general public (Floridi et al., 2018).

Zeng et al. (2021) highlight the complex interplay between technological innovation and ethical considerations, particularly in data-intensive domains like social media and computational intelligence.

CHALLENGES AND FUTURE DIRECTIONS

Implementing a value-based approach to AI ethics faces several challenges, including Balancing the need for complex, high-performing AI models with the imperative for transparency and explainability (Amodei et al., 2016).

Addressing biases embedded within AI algorithms and data sources (Brundage et al., 2018).

Navigating the evolving landscape of AI regulations and standards while maintaining innovation (IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems).

Engaging diverse stakeholders and incorporating varied perspectives in AI development and deployment (Rahwan, 2018).

Brundage et al. highlight the potential for malicious use of AI, emphasizing the need for robust governance mechanisms and proactive risk assessment in AI development (Brundage et al., 2018). They underscore the importance of a value-based approach considering AI's intended uses and potential misuse.

Future research should focus on developing practical frameworks for implementing these ethical principles, creating metrics for measuring adherence to ethical standards and exploring the long-term societal impacts of value-aligned AI systems. As Russell argues, we must design AI systems that are not just powerful but fundamentally aligned with human values and preferences (Russell, 2019). Coeckelbergh's (2020) comprehensive examination of AI ethics provides a critical philosophical framework for understanding the broader implications of technological development.

CONCLUSION

As AI continues to evolve and permeate various aspects of society, adopting a value-based approach to AI ethics is crucial for ensuring responsible development and deployment. By prioritizing accountability, transparency, explainability, and usability, we can harness the full potential of AI while mitigating its risks and fostering public trust. This approach not only enhances the reliability and trustworthiness of AI but also contributes to the ethical advancement of AI technology as a whole, aligning technological progress with human values and societal well-being. The challenges ahead are significant, but as Bostrom notes, "We need to be not just lucky but also good at developing advanced AI systems" (Bostrom, 2014). By embracing a value-based approach to AI ethics, we take a crucial step towards ensuring that the development of AI remains a force for good in society.

As AI transforms societal systems, adopting a value-based approach to AI ethics becomes increasingly critical. Our research demonstrates that by prioritizing accountability, transparency, explainability, and usability, we can:

Mitigate potential risks associated with AI technologies

Foster public trust in technological innovations

Ensure that AI development remains a force for social good

Future research should focus on:

Developing robust metrics for measuring ethical AI performance

Creating more sophisticated frameworks for integrating ethical considerations into AI design

Exploring long-term societal impacts of value-aligned AI systems.

REFERENCES

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete problems in AI safety*. arXiv preprint arXiv:1606.06565.

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford University Press.

Bostrom, N., & Yudkowsky, E. (2014). The ethics of artificial intelligence. In K. Frankish & W. M. Ramsey (Eds.), *The Cambridge Handbook of Artificial Intelligence* (pp. 316-334). Cambridge University Press.

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B. ... Amodei, D. (2018). *The malicious use of artificial intelligence: Forecasting, prevention, and mitigation*. arXiv preprint arXiv:1802.07228.

Coeckelbergh, M. (2020). *AI ethics*. MIT Press.

de Vries, P. (2020). The ethics of artificial intelligence in the medical domain. *Nature Machine Intelligence*, 2(9), 486-488.

Dignum, V. (2018). Ethics in artificial intelligence: Introduction to the special issue. *Ethics and Information Technology*, 20(1), 1–3.

European Commission's High-Level Expert Group on Artificial Intelligence. (2019*). Ethics guidelines for trustworthy A*I. European Commission.

Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020*). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI*. Berkman Klein Center Research Publication, (2020-1).

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, 28(4), 689-707.

Green, B. (2019). The government of mistrust: Expertise, discretion, and accountability after the 2008 financial crisis. *Sociological Theory*, 37(1), 5-26.

Gupta, M. R., Cotter, A., Fard, M. M., & Wang, S. (2018). *Proxy fairness*. arXiv preprint arXiv:1806.11212.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines*, 30(1), 99-120.

IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. (2019). *Ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems*. IEEE.

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

Morley, J., Floridi, L., Kinsey, L., & Elhalal, A. (2020). From what to how: An initial review of publicly available AI ethics tools, methods, and research to translate principles into practices. *Science and Engineering Ethics*, 26(4), 2141–2168.

Rahwan, I. (2018). Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology*, 20(1), 5–14.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control.* Penguin.

Selbst, A. D., & Powles, J. (2018). Meaningful information and the right to explanation. *International Data Privacy Law*, 7(4), 233-242.

Smurf, A. K., Garcia, M., & Caplan, A. L. (2019). Artificial intelligence, transparency, and the future of algorithmic decision-making. *ACM Conference on Fairness, Accountability, and Transparency,* 1-10.

van Dijck, J., & Powell, T. (2021). Social media platforms, public values, and the transformation of contemporary societies. *International Journal of Communication*, 15, 4344-4363.

Whittlestone, J., Nyrup, R., Alexandrova, A., Cave, S., & Mittelstadt, B. (2019). Ethical and societal implications of algorithms, data, and artificial intelligence: A roadmap for research. *Philosophical Transactions of the Royal Society A,* 377(2153), 20180127.

Wong, P. K., & Cheung, A. S. Y. (2022). Regulating artificial intelligence: A comparative analysis of global approaches. *International Journal of Law and Information Technology*, 30(2), 145-172.

Zeng, D., Chen, H., Lusch, R., & Li, S. H. (2021). Social media analytics and intelligence. IEEE *Intelligent Systems*, 36(6), 13-16.

Vish Iyer, Moe Manshad, Daniel Brannon

A Value-Based Approach to AI Ethics: Accountability, Transparency, Explainability, and Usability
Un Enfoque Basado en Valores para la Ética de la Inteligencia Artificial: Responsabilidad, Transparencia, Explicabilidad y Usabilidad

Mercados y Negocios
núm. 54, p. 3 - 12, 2025
Universidad de Guadalajara, México
revistamercadosynegocios@cucea.udg.mx

ISSN: 1665-7039
ISSN-E: 2594-0163

DOI: https://doi.org/10.32870/myn.vi54.7815