Towards Transparent Ethical AI: A Roadmap for Trustworthy Robotic Systems

Ahmad Farooq*[1] and Kamran Iqbal[2]

Abstract—**As artificial intelligence (AI) and robotics increasingly permeate society, ensuring the ethical behavior of these systems has become paramount. This position paper contends that transparency in AI decision-making processes is fundamental to developing trustworthy and ethically aligned robotic systems. We explore how transparency facilitates accountability, enables informed consent, and supports the debugging of ethical algorithms. The paper outlines technical, ethical, and practical challenges in implementing transparency and proposes novel approaches to enhance it, including standardized metrics, explainable AI techniques, and user-friendly interfaces. This paper introduces a framework that connects technical implementation with ethical considerations in robotic systems, focusing on the specific challenges of achieving transparency in dynamic, real-world contexts. We analyze how prioritizing transparency can impact public trust, regulatory policies, and avenues for future research. By positioning transparency as a fundamental element in ethical AI system design, we aim to add to the ongoing discussion on responsible AI and robotics, providing direction for future advancements in this vital field.**

Index Terms—**Ethical AI, Transparency, Explainable AI, Robotic Systems, Human-Robot Interaction**

I. Introduction

The swift progress in artificial intelligence (AI) and robotics has brought about an era of autonomous systems with complex decision-making capabilities. The rapid pace of AI advancements, including large language models and autonomous systems, has highlighted the need for transparent and ethical AI decision-making. These robotic systems are now part of many aspects of daily life—ranging from healthcare and transportation to manufacturing and security—thereby raising important ethical questions about their decision-making processes [1], [2].

Achieving transparency in AI involves more than just making the code accessible. It requires comprehending, interpreting, and explaining the logic behind a system's decisions and behaviors. In ethical robotics, transparency is not just a technical consideration but a fundamental ethical principle underpinning trust, accountability, and responsible innovation [3], [4]. However, current approaches to ethical AI in robotics often lack sufficient transparency, creating an opaque "black box" that hinders our ability to verify ethical compliance and address potential biases or errors.

This opacity in AI decision-making processes poses significant challenges. It impedes our ability to ensure accountability, obtain informed consent from users and stakeholders, and effectively debug and improve ethical algorithms [5], [6]. Moreover, the lack of transparency can erode public trust in robotic systems, potentially slowing their adoption and limiting their societal benefits.

This paper argues that transparency should be elevated to a fundamental principle in the development of ethical robotic systems. By illuminating the decision-making processes of AI-driven robots, we can build trust, facilitate meaningful human oversight, and ultimately create more ethically aligned robotic systems [7]. While acknowledging the technical, ethical, and practical hurdles in implementing transparency, this paper proposes innovative approaches to enhance it and discusses the far-reaching implications of this focus for the future of ethical AI in robotics.

The remainder of this paper is structured as follows: Section II presents the case for transparency in AI systems. Section III discusses the challenges in implementing transparency, while Section IV proposes approaches to enhance transparency in ethical AI decision-making. Section V examines the implications of prioritizing transparency and outlines future research directions. Finally, Section VI concludes the paper along with a call to action for the robotics and AI community.

II. The Case for Transparency

Lipton [8] defines transparency as comprising both model interpretability and post-hoc explanations. In this paper, we adopt a broad, layered perspective of transparency, ensuring it is both technically accessible and comprehensible to users. Transparency in robotic systems refers to offering clear and comprehensible insights into how the system makes decisions—covering all aspects from the data inputs to the algorithms applied and the reasoning behind the outcomes. This idea transcends basic technical openness, requiring accessibility and interpretability for a wide range of stakeholders, including users, developers, policymakers, and ethicists [9].

To determine if a system truly embodies transparency, we propose several criteria:

1) **Algorithmic Transparency:** The ability to inspect and understand the core algorithms and data processing techniques that are applied within the system.
2) **Functional Transparency:** Clear explanations of the system's functions, limitations, and the intended use cases.
3) **Operational Transparency:** Real-time analysis of the decision-making processes of the system during its

This work was not supported by any organization.

*Corresponding Author: Ahmad Farooq

[1] Ahmad Farooq is a Ph.D. Candidiate in the Electrical and Computer Engineering Department at the University of Arkansas at Little Rock, AR, 72204, USA afarooq@ualr.edu

[2] Kamran Iqbal is a Professor in the Electrical and Computer Engineering Department at the University of Arkansas at Little Rock, AR, 72204, USA kxiqbal@ualr.edu

operation.

4) **Ethical Transparency:** The disclosure of the ethical principles and considerations that are incorporated into the system's design and operation.

These criteria collectively provide a framework for evaluating and integrating transparency into AI and robotic systems, encompassing both technological and ethical aspects.

The importance of transparency in ethical AI is manifold:

1) **Enhancing Accountability:** Transparency enables the scrutiny of a system's actions and decisions. In situations involving significant risks—such as when robotic systems make decisions that could impact human health or safety, as in medical diagnostic tools or autonomous vehicles—the ability to trace these decisions becomes critical for assigning accountability and resolving issues [10], [11].
2) **Ensuring Informed Consent:** As robotic systems become more pervasive, people interacting with these systems have the right to understand how decisions affecting them are being made. This transparency is vital for maintaining human autonomy and privacy in interactions with AI [12], [13]. However, this right is not universally applicable and needs well-defined boundaries, as outlined in Section V.
3) **Assisting Algorithmic Debugging:** Transparency of AI system decision-making helps developers to effectively identify and address biases, errors, or unintended behaviors. Such improvements are fundamental for building reliable and ethically aligned AI systems [14], [15].
4) **Building Public Trust:** Greater transparency can lead to increased understanding and acceptance of robotic systems, thus accelerating their adoption in various sectors [16].

Despite the importance of transparency, many current robotic systems fall short in this regard. A significant number of AI algorithms, particularly those based on deep learning, are treated as "black boxes," which means their decision-making processes remain opaque [17]. This lack of transparency leads to numerous challenges, such as difficulties in detecting and correcting biases, challenges in verifying compliance with ethical and legal standards, reduced public trust, and obstacles in conducting effective ethical reviews and audits [18]–[20].

III. Challenges in Implementing Transparency

While the case for transparency in ethical AI decision-making is compelling, implementing it in practice presents several significant challenges. These can be broadly categorized into technical, ethical, and practical challenges, as summarized in Table I.

A. Technical Challenges

TABLE I: Challenges in Implementing Transparency in AI Systems

Category	**Challenge**	**Implications**
Technical	Complexity of AI algorithms	Difficulty in providing simple explanations
Ethical	Privacy concerns	Balancing transparency with data protection
Practical	User comprehension	Conveying complex information to diverse users

Modern AI systems, particularly those based on transformer models and deep learning, often involve intricate architectures with billions of parameters, making it difficult to provide simple, human-understandable explanations of their decision-making processes [21], [22]. There's often a perceived trade-off between the predictive power of an AI model and its interpretability; highly accurate models tend to be more complex and less transparent, while more interpretable models may sacrifice some degree of performance [23], [24].

However, Rudin and Radin [4] have challenged this perceived trade-off between transparency and model performance. The authors argue that interpretable models can achieve accuracy comparable to "black box" models, questioning the necessity of using opaque AI in high-stakes domains. This perspective highlights the importance of critically examining our assumptions about the relationship between model complexity and performance.

B. Challenges Specific to Robotic Systems

Robotic systems face unique challenges in ensuring transparency, predominantly as a result of their physical presence and their direct interactions with real-world environments:

1) **Real-time Decision Making:** Robotic systems often operate in highly dynamic environments, requiring continuous learning and adaptation. This constant evolution adds layers of complexity in achieving transparency for these systems [25].In robotics, the capacity for real-time decision-making and adaptation to evolving conditions is essential. Crafting clear, understandable explanations for actions taken in such unpredictable environments requires innovative strategies that can keep pace with the rapid changes the system undergoes.
2) **Multi-modal Interactions:** Robots integrate multiple sensors and actuators, leading to decision-making processes that are inherently complex and multi-modal. These multi-layered interactions are much harder to explain than those of purely data-driven AI systems.
3) **Safety-Critical Operations:** Many robotic systems operate in areas where safety cannot be compromised. In these cases, transparency must be delicately balanced with the need for rapid, dependable performance, as revealing too much may jeopardize system responsiveness or reliability.
4) **Human-Robot Interaction:** Effectively conveying a robot's goals and decision-making processes in real-time, in a manner comprehensible to humans, continues to pose a significant challenge. This necessitates the creation of intuitive, user-friendly communication

mechanisms that effectively communicate the robot's activities and intents.

C. Ethical Challenges

Although transparency is essential, it must be weighed against the need to safeguard sensitive information and protect individual privacy. In certain situations, full transparency might expose proprietary data or compromise user confidentiality [26], [27]. Providing detailed explanations of a system's decision-making process could also be misused by malicious actors to manipulate or exploit the system [28].

Additionally, efforts to make complex systems easier to understand carry the risk of oversimplification, which may result in misunderstandings or misplaced confidence [29]. Finding the right balance between providing enough detail and ensuring comprehensibility remains a significant challenge, especially when addressing the varied needs of different stakeholders, ranging from technical experts to end-users.

D. Practical Challenges

Achieving transparency on a technical level is only part of the challenge; effectively communicating this information in a way that is clear and meaningful to users with diverse levels of technical expertise remains difficult [30]. In dynamic settings where robots need to make split-second decisions, providing real-time explanations without compromising system performance is particularly challenging [31].

Implementing transparency initiatives may also demand extra computational resources, which can affect the cost-effectiveness and efficiency of robotic systems [32]. Moreover, promoting transparency in ethical AI involves knowledge from multiple disciplines, including computer science, ethics, cognitive science, and human-computer interaction, making it a complex, interdisciplinary issue [33].

IV. Proposed Approaches to Enhance Transparency

To address the challenges outlined in the previous section and move towards more transparent ethical AI decision-making in robotic systems, we propose several approaches. Table II provides a comparison of these transparency approaches, highlighting their advantages and challenges.

TABLE II: Comparison of Transparency Approaches

Approach	Advantages	Challenges
Standardized Metrics	Quantifiable and comparable across systems	Difficult to standardize across diverse AI applications
XAI Techniques	Provides insights into complex models	May reduce model performance
User-Friendly Interfaces	Improves user understanding and trust	Requires significant design effort
Transparency-by-Design	Proactive approach, Integrates ethics early	May slow initial development process

A. Targeted Transparency Approaches for Different Stakeholders

The proposed transparency measures cater to both system designers and end-users. For system designers, the focus is on providing in-depth metrics and insights for debugging and improvement. This includes detailed information about the AI models, training data, and decision-making processes. For end-users, emphasis is placed on providing intuitive, user-friendly explanations that convey the system's capabilities, limitations, and basic decision-making rationale without requiring technical expertise.

B. Developing Standardized Transparency Metrics

Creating a comprehensive transparency index that quantifies the level of transparency in a robotic system, considering factors such as explainability, interpretability, and accessibility of information, is crucial [34]. This should be complemented by collaborating with standards organizations to create widely accepted transparency benchmarks for different types of robotic systems and applications [35]. Implementing a framework for periodic assessments of robotic systems against these standardized metrics can ensure ongoing compliance and improvement [36].

C. Incorporating Explainable AI (XAI) Techniques

Where possible, AI models that are intrinsically more interpretable, such as decision trees or rule-based systems, should be used for critical ethical decision-making components [37]. For complex models like deep neural networks, techniques such as Local Interpretable Model-agnostic Explanations (LIME) or SHapley Additive exPlanations (SHAP) can give insights into decision-making processes [38]. Developing hybrid systems that combine the power of complex AI models with more transparent, rule-based systems for ethical decision-making is another promising approach [39].

D. Creating User-Friendly Interfaces

Designing interfaces that provide explanations through various modalities (e.g., visual, textual, auditory) can cater to different user preferences and cognitive styles [40]. Implementing AI-driven interfaces that adjust the level and complexity of explanations based on the user's expertise and context can enhance understanding [41]. Developing tools that enable users to explore the decision-making process interactively, allowing them to ask questions and receive relevant explanations, can further improve transparency [42].

E. Establishing Transparency Requirements in Design

Integrating transparency considerations from the earliest stages of robotic system design, making it a fundamental requirement rather than an afterthought, is essential [43]. Conducting thorough ethical impact assessments during the design phase can help identify potential ethical issues and transparency needs [44]. Involving diverse stakeholders, including ethicists, end-users, and policymakers in the design process can ensure transparency measures meet varied needs and expectations [45].

V. Implications and Future Directions

Prioritizing transparency in ethical AI decision-making for robotic systems has far-reaching implications and opens up several avenues for future research and development. Table III outlines key research areas and associated questions for future work in AI transparency.

TABLE III: Future Research Directions in AI Transparency

Research Area	Key Questions and Objectives
Cognitive Models of Explanation	How do humans process and understand AI explanations?
Multi-Agent Transparency	How can transparency be maintained in systems with multiple AI agents?
Long-term Impact Studies	What are the long-term effects of increased AI transparency on public trust?
AI Literacy Programs	How can we effectively educate the public about AI decision-making?
Domain-Specific Transparency	What are the unique transparency needs in healthcare, autonomous vehicles, etc.?

A. Impact on Public Trust and Adoption

Increased transparency can lead to greater understanding and acceptance of robotic systems, potentially accelerating their adoption in various sectors [46]. Transparent systems may facilitate more effective teamwork between humans and robots, as humans can better understand and predict robot behavior [47]. Moreover, transparency metrics could enable consumers to make more informed decisions about the robotic products and services they use, driving market demand for ethical AI [48].

B. Implications for Regulatory Frameworks

Transparent AI systems provide policymakers with clearer insights into robot decision-making, enabling more informed and effective regulation [49]. The push for transparency could drive international efforts to standardize ethical AI practices, similar to existing standards in other technological domains [50]. Enhanced transparency may contribute to creating more sophisticated legal frameworks for determining liability in situations involving autonomous robotic systems [51].

C. Ethical Reflections on the Right to Explanation

The ethical duty to provide transparency must take into account the right to explanation, similar to human decision-making processes. However, it is essential to critically evaluate the extent and limits of this right in interactions between humans and AI. Unlike interactions between humans, where such rights may vary by context, interactions between humans and robots often require a higher level of clarity, especially in safety-critical situations.

The right to explanation is particularly vital in fields such as healthcare or criminal justice, where AI decisions significantly impact individuals' lives. Nevertheless, this right may not always be applicable or absolute. Key factors that influence the scope of this right include:

- The potential effect of the decision on human lives and well-being
- The complexity involved in the decision-making process
- The urgency of the decision
- Considerations related to privacy and security
- The technical feasibility of delivering an understandable explanation

Continued research and ethical discussions are necessary to establish clear guidelines regarding when and how this right should be implemented in different human-AI interaction scenarios.

D. Future Research Directions

Several key areas for future research emerge from our analysis:

1) **Cognitive Models of Explanation:** Further research into how humans process and understand explanations can inform the development of more effective transparency mechanisms [52].
2) **Multi-Agent Transparency:** Exploring how to maintain transparency in complex scenarios involving multiple interacting robotic agents is another crucial area of study [53].
3) **Long-term Impact Studies:** Longitudinal studies on how increased transparency affects public perception, trust, and interaction with robotic systems over time are needed [54].
4) **AI Literacy Programs:** Developing educational programs to improve public understanding of AI decision-making processes and ethical considerations is essential [55].
5) **Domain-Specific Transparency:** Investigating transparency requirements and solutions for specific applications of robotic systems, particularly in areas like healthcare, autonomous vehicles, and smart cities, will be crucial for the ethical development of these technologies [56].

VI. Conclusion

This paper has advocated for transparency as a fundamental ethical principle in AI decision-making for robotic systems, underpinning trust, accountability, and responsible innovation. We've proposed a comprehensive framework addressing technical and ethical aspects, including standardized metrics, explainable AI techniques, user-friendly interfaces, and design-phase transparency requirements. The complex relationship between transparency and trust necessitates coupling transparency efforts with broader ethical considerations and stakeholder dialogue.

We urge the robotics and AI community to prioritize transparency through interdisciplinary collaborations. Future work should focus on developing standardized metrics, advancing explainable AI, improving human-robot interaction interfaces, and researching long-term impacts on public trust and AI adoption. By elevating transparency to a core principle, we can encourage responsible technological advancement, realizing AI and robotics' potential to improve human life while ensuring accountability and alignment with human values.

References

[1] A. F. Winfield and M. Jirotka, "Ethical governance is essential to building trust in robotics and artificial intelligence systems," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, no. 2133, p. 20180085, 2018.

[2] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. G.-L. Garcia, D. Molina, R. Benjamins, R. Chatila *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[3] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, B. Schafer, P. Valcke, and E. Vayena, "Ai4people—an ethical framework for a good ai society: Opportunities, risks, principles, and recommendations," *Minds and Machines*, vol. 28, no. 4, pp. 689–707, 2018.

[4] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[5] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[6] S. Cave and K. Dihal, "Hopes and fears for intelligent machines in fiction and reality," *Nature Machine Intelligence*, vol. 1, no. 2, pp. 74–78, 2019.

[7] J. Bryson and A. Winfield, "Standardizing ethical design for artificial intelligence and autonomous systems," *Computer*, vol. 50, no. 5, pp. 116–119, 2017.

[8] Z. C. Lipton, "The mythos of model interpretability," *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[9] S. Wachter, B. Mittelstadt, and L. Floridi, "Transparent, explainable, and accountable ai for robotics," *Science Robotics*, vol. 2, no. 6, 2017.

[10] V. Dignum, *Responsible artificial intelligence: How to develop and use AI in a responsible way*. Springer Nature, 2019.

[11] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan, "The moral machine experiment," *Nature*, vol. 563, no. 7729, pp. 59–64, 2018.

[12] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature Machine Intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[13] M. Tegmark, *Life 3.0: Being human in the age of artificial intelligence*. Knopf, 2017.

[14] D. Gunning and D. W. Aha, "Darpa's explainable artificial intelligence (xai) program," *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019.

[15] S. Russell, *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019.

[16] K. A. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Human Factors*, vol. 57, no. 3, pp. 407–434, 2015.

[17] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, Eds., *Explainable AI: Interpreting, explaining and visualizing deep learning*. Springer Nature, 2019, vol. 11700.

[18] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[19] J. Fjeld, N. Achten, H. Hilligoss, A. Nagy, and M. Srikumar, "Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai," *Berkman Klein Center Research Publication*, no. 2020-1, 2020.

[20] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, "Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*. ACM, 2020, pp. 33–44.

[21] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[22] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2018, pp. 80–89.

[23] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (xai)," *IEEE Access*, vol. 6, pp. 52 138–52 160, 2018.

[24] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "Xai—explainable artificial intelligence," *Science Robotics*, vol. 4, no. 37, 2019.

[25] L. Deng, "Artificial intelligence in the rising wave of deep learning: The historical path and future outlook [perspectives]," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 180–177, 2018.

[26] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the gdpr," *Harvard Journal of Law & Technology*, vol. 31, p. 841, 2017.

[27] B. Lepri, N. Oliver, E. Letouzé, A. Pentland, and P. Vinck, "Fair, transparent, and accountable algorithmic decision-making processes," *Philosophy & Technology*, vol. 31, no. 4, pp. 611–627, 2018.

[28] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in ai safety," *arXiv preprint arXiv:1606.06565*, 2016.

[29] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[30] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable ai: Challenges and prospects," *arXiv preprint arXiv:1812.04608*, 2018.

[31] S. Anjomshoae, A. Najjar, D. Calvaresi, and K. Främling, "Explainable agents and robots: Results from a systematic literature review," in *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS 2019)*. International Foundation for Autonomous Agents and Multiagent Systems, 2019, pp. 1078–1088.

[32] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you? explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1135–1144.

[33] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, S. Schieber, J. Waldo, D. Weinberger, and A. Wood, "Accountability of ai under the law: The role of explanation," *arXiv preprint arXiv:1711.01134*, 2017.

[34] S. Mohseni, N. Zarei, and E. D. Ragan, "A multidisciplinary survey and framework for design and evaluation of explainable ai systems," *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 11, no. 3–4, pp. 1–45, 2021.

[35] M. Brundage, S. Avin, J. Wang, H. Belfield, G. Krueger, G. Hadfield, H. Khlaaf, J. Yang, H. Toner, M. Anderljung *et al.*, "Toward trustworthy ai development: Mechanisms for supporting verifiable claims," *arXiv preprint arXiv:2004.07213*, 2020.

[36] I. D. Raji and J. Buolamwini, "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 2019, pp. 429–435.

[37] C. Rudin and J. Radin, "Why are we using black box models in ai when we don't need to? a lesson from an explainable ai competition," *Harvard Data Science Review*, vol. 1, no. 2, 2019.

[38] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 4765–4774.

[39] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE, 2018, pp. 0210–0215.

[40] A. Abdul, J. Vermeulen, D. Wang, B. C. Y. Lim, and M. Kankanhalli, "Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 2018, pp. 1–18.

[41] D. Wang, Q. Yang, A. Abdul, and B. C. Y. Lim, "Designing theory-driven user-centric explainable ai," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 2019, pp. 1–15.

[42] D. S. Weld and G. Bansal, "The challenge of crafting intelligible intelligence," *Communications of the ACM*, vol. 62, no. 6, pp. 70–79, 2019.

[43] V. Dignum, M. Baldoni, C. Baroglio, M. Caon, R. Chatila, L. Dennis *et al.*, "Ethics by design: Necessity or curse?" in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 2018, pp. 60–66.

[44] D. Wright and M. Friedewald, "Integrating privacy and ethical impact assessments," *Science and Public Policy*, vol. 40, no. 6, pp. 755–766, 2013.

[45] T. Hagendorff, "The ethics of ai ethics: An evaluation of guidelines," *Minds and Machines*, vol. 30, no. 1, pp. 99–120, 2020.

[46] E. J. de Visser, R. Pak, and T. H. Shaw, "From 'automation' to 'autonomy': The importance of trust repair in human–machine interaction," *Ergonomics*, vol. 61, no. 10, pp. 1409–1427, 2018.

[47] B. Shneiderman, "Human-centered artificial intelligence: Reliable, safe & trustworthy," *International Journal of Human–Computer Interaction*, vol. 36, no. 6, pp. 495–504, 2020.

[48] L. Floridi, "Establishing the rules for building trustworthy ai," *Nature Machine Intelligence*, vol. 1, no. 6, pp. 261–262, 2019.

[49] C. Cath, S. Wachter, B. Mittelstadt, M. Taddeo, and L. Floridi, "Artificial intelligence and the 'good society': The us, eu, and uk approach," *Science and Engineering Ethics*, vol. 24, no. 2, pp. 505–528, 2018.

[50] A. Theodorou and V. Dignum, "Towards ethical and socio-legal governance in ai," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 10–12, 2020.

[51] J. K. Kingston, "Artificial intelligence and legal liability," in *Research and Development in Intelligent Systems XXXIII*. Springer, Cham, 2018, pp. 269–279.

[52] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.

[53] A. Dafoe, E. Hughes, Y. Bachrach, T. Collins, K. R. McKee, J. Z. Leibo, K. Larson, and T. Graepel, "Open problems in cooperative ai," *arXiv preprint arXiv:2012.08630*, 2020.

[54] S. Cave, K. Coughlan, and K. Dihal, "Scary robots: Examining public responses to ai," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 2019, pp. 331–337.

[55] D. Long and B. Magerko, "What is ai literacy? competencies and design considerations," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 2020, pp. 1–16.

[56] C. Trocin, P. Mikalef, Z. Papamitsiou, and K. Conboy, "Responsible ai for digital health: a synthesis and a research agenda," *Information Systems Frontiers*, vol. 25, no. 6, pp. 2139–2157, 2023.