Ethics, Accountability, and the Pursuit of Responsible AI

As artificial intelligence (AI) systems become increasingly common in healthcare, finance, public administration, and other sectors, their impact on society continues to grow rapidly. While AI's continued adoption and deployment promises innovation and improved outcomes, it also raises ethical concerns related to fairness, transparency, and accountability.[1] Issues such as algorithmic bias, lack of interpretability, and cross-border regulatory challenges highlight the urgent need to redefine accountability.

Accountability is critical and encompasses clarifying roles and responsibilities to ensure that harms are addressed and mitigated through effective oversight. However, many of the current accountability models are not equipped to manage the decentralized and autonomous nature of emerging technologies.[2] Addressing accountability in AI deployment is essential to safeguarding integrity and societal well-being. This will require increased transparency within explainable AI and strengthened ethical governance while ensuring diverse oversight and public empowerment via widespread AI education. Addressing the shortcomings of traditional accountability frameworks in the face of rapidly evolving AI technology is nothing short of critical.

Ethical Considerations

Four key ethical challenges of AI consistently stand out due to their potential to impact society.

Bias and Discrimination

Algorithms can unintentionally encode and perpetuate biases from training data. This occurred in France in 2024, when a human rights organization filed a legal challenge against the French government's use of algorithms to detect welfare payment errors, claiming the systems discriminated against disabled individuals and single mothers.[3] Ensuring unbiased and impartial implementation of AI requires the thorough consideration of data quality and ethical implications through structured oversight and accountability. A class-action lawsuit was settled against SafeRent Solutions after its algorithm disproportionately denied housing to Black applicants, resulting in a payment of more than US$2.2 million and forced revision to their screening processes.[4] Organizational practices must incorporate rigorous validation to identify and mitigate bias throughout the entire AI life cycle, ensuring ethical and equitable technology implementations.

JOSHUA SCARPINO | D.SC

Is the vice president of information security at TrustEngine. He leads the IT operations and security and compliance teams and is also responsible for developing and managing the organization's responsible AI program. Scarpino is also the chief executive officer and founder of Assessed Intelligence, a cybersecurity and AI research consultancy firm. He has more than two decades of experience working in various US Department of Defense (DOD) roles, leading security operations for Fortune 500 companies, and enhancing critical controls at financial and manufacturing organizations. He has also led, scaled, and audited multiple security and compliance programs. Scarpino is a Fellow with ForHumanity; a member of the US National Institute of Technology (NIST) AI Safety Working Group; a Cybersecurity Professor at Stark State College (Ohio, USA); and a member of the Ohio Cyber Reserves (USA).

Explainability

Non-interpretable or difficult-to-interpret systems often hinder appropriate oversight and risk management activities. This is a consistent problem and has impacted organizations such as the UK Department for Work and Pensions. In 2024, the department faced criticism for its AI system used to detect citizen benefit fraud. Analysis revealed biases against individuals based on age, disability, marital status, and nationality, which raised concerns about the system's transparency and fairness.[5] Ensuring the explainability of AI systems is foundational to understanding and assigning accountability. Explainable AI (XAI) enhances transparency by enabling stakeholders to understand the reasoning behind automated decisions. Interpretability improves trust and facilitates alignment with regulatory compliance initiatives; this is foundational to XAI.[6] Research conducted by Anthropic and Redwood Research in 2025 showed that AI models, such as Claude, have the potential to strategically deceive their human creators. This highlights the challenges in aligning AI systems with human values and the difficulties surrounding the explainability of their decision-making processes.[7]

Organizational practices must incorporate rigorous validation to identify and mitigate bias throughout the entire AI life cycle, ensuring ethical and equitable technology implementations.

Techniques can be employed to combat these challenges, including post-process explanations and inherently interpretable architectures that help support comprehension.[8] To accomplish better interpretability in AI systems, they can be implemented using post-process explanations, where explanations for model outputs are generated after the predictions have been made.[9] Another approach mentioned is employing inherently interpretable architectures, where interpretability is integrated directly into the model's structure and design. This can help support immediate comprehension of how the model arrives at specific decisions.[10] Post-process explanations and inherently interpretable architectures can increase transparency and trust in AI. Selecting the most suitable method depends on the system design, application, ethical considerations, and specific requirements.[11] It is imperative for organizations to clearly understand the decisions made by AI; responsibility for these decisions must be held by a real person while maintaining explainability and repeatability.[12]

Organizations must prioritize explainability at every stage of the AI life cycle, from model selection through deployment and monitoring. Transparency should focus on technical details and ensure that decision-making processes align with ethical principles and societal norms. This requires cross-functional collaboration involving individuals from varied perspectives and experiences.[13]

Autonomy and Misuse

Autonomous systems can operate in a manner that deviates from the designed intent, leading to unpredictable outcomes for society. In 2024, the US House Subcommittee report raised alarms about the federal government's development and use of AI tools for mass monitoring and potential censorship. They highlighted the risk of autonomous systems being misused to suppress dissent and infringe on free speech at scale.[14] As organizations increasingly adopt AI, concerns regarding system autonomy and the potential misuse of these technologies have become critical. Implementing autonomous decision-making systems introduces risk associated with reduced human oversight, potentially leading to unintended consequences or misuse. Organizations must clearly define roles and responsibilities to maintain accountability, ensuring any ethical decisions that remain have human oversight. Studies have revealed that AI systems, including OpenAI's o1 and Anthropic's Claude 3.5 Sonnet, had the potential to show deceptive behaviors. This included hiding their capabilities and objectives from humans so that they could achieve specific goals.[15] Effective management of AI autonomy includes establishing clear governance frameworks to prevent misuse or harmful outcomes, thus protecting both organizational interests and individual rights.

Accountability

Accountability involves the obligation to explain, justify, and take responsibility for technological

actions and decisions. In most industries, accountability is straightforward, with clear expectations. This can be seen in sectors such as aviation. When an airplane accident occurs, clear regulatory guidelines outline the responsibilities of pilots, airlines, and manufacturers. In the case of the Boeing 737 MAX plane crashes that occurred in 2018 and 2019, investigations clearly identified specific lapses in Boeing's design process, resulting in direct accountability, extensive scrutiny, regulatory intervention, and substantial penalties.[16] In AI, accountability is often diffused, as systems operate autonomously and involve multiple stakeholders.[17] The lines between accountability and responsibility have increasingly overlapped in society.[18] Consider self-driving vehicles that rely on networks of sensors, algorithms, and machine learning models. When accidents occur, it is unclear whether responsibility lies with the manufacturer, developer, user, or data provider.[19]

Adopting structured frameworks that embed accountability into the AI life cycle is essential to strengthening accountability. The accountability by design approach ensures that accountability principles are incorporated at each development stage: design, deployment, and post-deployment monitoring.[20] As regulations evolve, algorithmic accountability reporting will increasingly be mandated for organizations deploying high-stakes AI systems. These reports could include information about the system's intended purpose, design choices, datasets used, and potential risk identified during development.[21]

Unfortunately, establishing accountability is not without its challenges:

- **Complexity and opacity**—The complexity of AI systems creates significant barriers to accountability. Many AI models, particularly those using deep learning, function as black box systems where even developers can struggle to understand their decision-making processes.[22] Lack of transparency raises concerns about the deployed system's fairness and reliability.
- **Distributed responsibility**—Emerging technologies often involve multiple stakeholders, including system developers, data providers, users, and regulators. This distributed nature complicates the assignment of liability for a given system.
- **Regulatory lag and cross-jurisdictional challenges**—Technological innovation often outpaces regulatory development, creating accountability gaps. Moreover, AI systems frequently operate across national borders, making it challenging to enforce consistent accountability standards. Differences in data protection laws, such as the EU's General Data Protection Regulation (GDPR)[23] and the sectoral approach of the United States, make these challenges more difficult to overcome.

Organizations must clearly define roles and responsibilities to maintain accountability, ensuring any ethical decisions that remain have human oversight.

Developing Explainable and Transparent AI

Accountability starts by building explainability into AI systems. This is essential to interpretability and fostering trust while ensuring regulatory compliance.[24] Techniques such as transparent and interpretable model architecture can make AI systems easier to understand, while strong ethical governance is paramount throughout the AI life cycle. Frameworks should be grounded in core principles such as fairness, transparency, and accountability.[25]

Ethical Governance and Oversight

Ethical considerations must be embedded throughout the entire life cycle of all high-risk AI systems, from design and development through deployment and monitoring.[26] The EU's Ethics Guidelines for Trustworthy AI provides a model that can be leveraged, focusing on core principles.[27] While independent oversight bodies can monitor compliance and address ethical concerns proactively, governance frameworks are essential for ensuring that AI aligns with human values and societal goals.[28]

The EU's Ethics Guidelines for Trustworthy AI provides a comprehensive framework that emphasizes human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental well-being, and accountability.[29] These principles offer a roadmap for developing and deploying AI systems that respect fundamental rights and promote human flourishing.

Achieving accountability in AI requires a multifaceted approach to stakeholder engagement, moving beyond consultation toward active participation and shared responsibility.

Independent oversight mechanisms will continue to play a crucial role in monitoring AI development and deployment, ensuring compliance with ethical standards, and proactively addressing potential harm. Such bodies can provide expert guidance, conduct audits, investigate complaints, and recommend policy initiatives to mitigate risk and promote responsible innovation.[30]

AI Accountability Solutions

Achieving accountability in AI requires a multifaceted approach to stakeholder engagement, moving beyond consultation toward active participation and shared responsibility. Multi-stakeholder feedback is critical when deploying high-risk AI systems. This involves recognizing various interests and perspectives, including the public, civil society organizations, academia, and industry, and integrating their ideas into AI governance.[31]

Empowerment Through Transparency, Education, and Collaborative Governance

Holding AI developers and deployers accountable helps drive transparency regarding AI systems and forces evaluations of their potential impacts.[32] This includes providing easily accessible information about how a given AI system works, the data used to train it, and the possible biases the system may exhibit due to the design process. Promoting literacy in AI and increased awareness through educational initiatives can enable individuals to engage critically with AI technologies, gain a depth of understanding of their rights, and participate meaningfully in public discourse.[33]

Shifting from a top-down approach to a participatory design process can ensure that AI systems are developed and deployed in an inclusive manner reflective of diverse societal values.[34] This involves actively soliciting multi-stakeholder feedback throughout all stages of the AI life cycle, from problem definition and data collection to model development evaluation and deployment. This empowers communities to have appropriate representation for technologies that can affect their lives, fostering a sense of ownership and ensuring equitable AI outcomes.

Establishing effective accountability requires collaborative governance that unites diverse stakeholders so they can deliberate on AI policy and practice.[35] Inclusivity is critical and involves creating stakeholder initiatives, advisory boards, and ethical review committees that include government, industry, academia, and representatives across society. Collaborative approaches facilitate dialogue, foster consensus-building, and ensure that accountability measures are grounded in diverse and various perspectives. Inclusivity minimizes the potential for disparate impacts before development and deployment.

Mechanisms for Redress

Ensuring accountability also requires establishing mechanisms for addressing issues and redressing the harm caused by AI systems.[36] This can involve creating independent oversight bodies such as the European Artificial Intelligence Board (EAIB) proposed under the EU AI Act, with the authority to investigate complaints, conduct audits, and impose sanctions for violations of ethical guidelines or legal frameworks.

Fostering a culture of responsible disclosure and whistleblowing can help encourage the identification and mitigation of potential AI harm. Additionally, validation can be accomplished by leveraging reporting channels; this offers an opportunity to understand the possible risk of a given system's decisions. Accepting the judgment from an AI system should not ever be seen as the last resort for impacted individuals. All technologies that leverage AI with potential disparate impacts should have clearly defined processes explaining how impacted users can challenge the decisions made. Additionally, as a foundational requirement, decisions from a high-impact AI system must be owned by a physical or legal person and be explainable, repeatable, and based on explicit scientific theories.[37]

Conclusion

Accountability is a foundational requirement for the ethical and effective deployment of AI and ensures that individuals are held responsible for the systems they deploy. As AI becomes increasingly woven into the fabric of daily life, a crucial question

LOOKING FOR MORE?

- Read *Artificial Intelligence: A Primer on Machine Learning, Deep Learning, and Neural Networks. https://www.isaca.org/ai-primer*
- Learn more about, discuss, and collaborate on emerging technologies in ISACA's Online Forums. *https://engage.isaca.org/onlineforum*

is posed: How do we ensure that individuals remain accountable to the systems they develop and their impact on society? The challenges of AI accountability highlight the need for transparency, ethical governance, education, and a human-centered approach to deployment. Traditional accountability models struggle to keep pace with AI's rapid evolution. To bridge this gap, core strategies include:

- **Demystify AI**—XAI is key to fostering trust and understanding. We must continue to empower human oversight and ensure responsible use by shedding light on AI's decision-making processes.
- **Embed ethics**—AI should not operate in an ethical vacuum where it is only controlled by a curated list of AI experts or technologists. Diverse perspectives should be considered, and fairness, transparency, and human well-being principles must be embedded into AI systems.
- **Strengthen oversight**—Independent bodies are crucial for monitoring AI, validating compliance with defined standards, and proactively addressing potential harm.
- **Empower through education**—Knowledge is power. Through the promotion of digital literacy and AI education and awareness, citizens and practitioners can be equipped to engage critically with AI, hold developers accountable, and ensure that varied perspectives are considered. AI accountability is a shared responsibility. Collaborative governance structures will bring together diverse voices to shape the future of AI.

By addressing these challenges, stakeholders can continue to foster trust and ensure that emerging technologies align with societal values. AI holds vast potential to improve the lives of many people across the world, but only if it remains aligned with society's values. By embracing a proactive and human-centered approach to AI governance, its power can be harnessed for good while safeguarding those it could disparately impact.

Endnotes

1 Scarpino, J.; "Evaluating Ethical Challenges in AI and ML," *ISACA® Journal,* vol. 4, 2022, *https://www.isaca.org/archives*; Santoni de Sio, F.; Mecacci, G.; "Four Responsibility Gaps With Artificial Intelligence: Why They Matter and How to Address Them," *Philosophy & Technology*, vol. 34, 2021, p. 1057–1084, *https://doi.org/10.1007/s13347-021-00450-x*

2 Scarpino; "Evaluating Ethical Challenges"

3 Meaker, M.; "Algorithms Policed Welfare Systems for Years. Now They're Under Fire for Bias," Wired, 16 October 2024, *https://www.wired.com/story/algorithms-policed-welfare-systems-for-years-now-theyre-under-fire-for-bias/*

4 Bedayn, J.; "Class Action Lawsuit on AI-Related Discrimination Reaches Final Settlement," Associated Press, 20 November 2024, *https://apnews.com/article/artificial-intelligence-ai-lawsuit-discrimination-bias-1bc785c24a1b88bd425a8fa367ab2b23*

5 Booth, R.; "Revealed: Bias Found in AI System Used to Detect UK Benefits Fraud," *The Guardian*, 6 December 2024, *https://www.theguardian.com/society/2024/dec/06/revealed-bias-found-in-ai-system-used-to-detect-uk-benefits*

6 Doshi-Velez, F.; Kim, B.; "Towards a Rigorous Science of Interpretable Machine Learning," 2017, arXiv, *https://doi.org/10.48550/arXiv.1702.08608*

7 Perrigo, B.; "Exclusive: New Research Shows AI Strategically Lying," Time, 2025, *https://time.com/7202784/ai-research-strategic-lying/*

8 Scarpino; "Evaluating Ethical Challenges"

9 Ribeiro, M. T.; Singh, S.; et al.; "Why Should I Trust You?: Explaining the Predictions of Any Classifier," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 1135–1144, 2016, *https://doi.org/10.1145/2939672.2939778*; Lundberg, S. M.; Lee, S.; "A Unified Approach to Interpreting Model Predictions," arXiv, 2017, *https://doi.org/10.48550/arXiv.1705.07874*

10 Rudin, C.; "Stop Explaining Black Box Models for High Stakes Decisions and Use Interpretable Models Instead," *Nature Machine Intelligence*, vol. 1, iss. 5, 2019, p. 206–215, *https://doi.org/10.1038/s42256-019-0048-x*

11 Lipton, Z.; "The Mythos of Model Interpretability," *Communications of the ACM*, vol. 61, iss. 10, 2018, p. 36–43, *https://doi.org/10.1145/3233231; Rudin;* "Stop Explaining Black Box"

12 Muller, H.; Mayrhofer, M. T.; et al.; "The Ten Commandments of Ethical Medical AI," *Computer*, vol. 54, iss. 7, 2021, p. 119–123, *https://doi.org/10.1109/MC.2021.3074263*

13 Scarpino, J.; *An Exploratory Study: Implications of Machine Learning and Artificial Intelligence in Risk Management*, Marymount University, Arlington, Virginia, USA, 2022, *https://www.proquest.com/docview/2731478908*

14 Lifson, A.; "Threat of 'Mass' AI-Powered Government Censorship is on the Rise, Ominous House Panel Report Shows," *New York Post*, 18 December 2024, *https://nypost.com/2024/12/18/us-news/threat-of-mass-ai-powered-government-censorship-is-on-the-rise-ominous-house-panel-report-shows/*

15 Perrigo, B.; "New Tests Reveal AI's Capacity for Deception," *Time*, 25 January 2025, *https://time.com/7202312/new-tests-reveal-ai-capacity-for-deception/*

16 Seattle Times Staff, "Boeing 737 MAX Timeline: Troubled History, Uncertain Future," The Seattle Times, 14 January 2020, *https://www.seattletimes.com/business/boeing-aerospace/boeing-737-max-timeline-troubled-history-uncertain-future/*

17 Floridi, L.; Cowls, J.; et al.; "AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations," *Minds & Machine*s, vol. 28, 2018, p. 689–707, *https://doi.org/10.1007/s11023-018-9482-5*

18 Bovens, M.; "Two Concepts of Accountability: Accountability as a Virtue and as a Mechanism," *West European Politics*, vol. 33, iss. 5, 2010, p. 946–967, *https://doi.org/10.1080/01402382.2010.486119*

19 Gurney, J. K.; "Sue My Car Not Me: Product Liability and Accidents Involving Autonomous Vehicles," *University of Illinois Journal of Law, Technology & Policy,* 2017; Scarpino; An Exploratory Study

20 Novelli, C.; Taddeo, M.; et al.; "Accountability in Artificial Intelligence: What It Is and How It Works," AI & Society, vol. 39, 2024, p. 1871–1882, *https://doi.org/10.1007/s00146-023-01635-y*

21 Floridi, L.; Cowls, J.; "A Unified Framework of Five Principles for AI in Society," *Harvard Data Science Review*, vol. 1, iss. 1, 2019, *https://doi.org/10.1162/99608f92.8cd550d1*

22 Doshi-Velez; Kim; "Towards a Rigorous Science"

23 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation [GDPR])

24 Scarpino; *An Exploratory Study*; Doshi-Velez; Kim; "Towards a Rigorous Science"

25 Gilpin, L. H.; Bau, D.; et al.; "Explaining Explanations: An Overview of Interpretability of Machine Learning," *Proceedings of the 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 2018, p. 80-89, *https://doi.org/10.1109/DSAA.2018.00018*

26 Jobin, A.; Ienca, M.; et al.; "The Global Landscape of AI Ethics Guidelines," *Nature Machine Intelligence*, vol. 1, iss. 9, 2019, p. 389-399, *http://dx.doi.org/10.1038/s42256-019-0088-2; Scarpino; An Exploratory Study; Doshi-Velez; Kim; "Towards a Rigorous Science"*

27 European Commission, *Ethics Guidelines for Trustworthy AI, European Union*, 2019, *https://op.europa.eu/en/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1*

28 Mittelstadt, B. D.; Allo, P.; et al.; "The Ethics of Algorithms: Mapping the Debate," *Big Data & Society*, vol. 3, iss. 2, 2016, *https://doi.org/10.1177/2053951716679679*

29 European Commission, *Ethics Guidelines*

30 Casey, A.; Farhangi, A.; et al.; "Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise," *California Law Review*, vol. 108, iss. 4, 2020, p. 885-935; O'Neil, C.; Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, Crown, USA, 2016

31 Floridi; Cowls; "A Unified Framework"

32 Ananny, M.; Crawford, K.; "Seeing Without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability," *New Media & Society*, vol. 20, iss. 3, 2018, p. 973-989, *https://doi.org/10.1177/1461444816676645*

33 UNESCO, *Recommendation on the Ethics of Artificial Intelligence*, 2021, *https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence*; Scarpino; An Exploratory Study

34 Jasanoff, S.; *The Ethics of Invention: Technology and the Human Future*, WW Norton & Company, USA, 2016

35 Whittlestone, J.; Nyrup, R.; et al.; "The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions," *Proceedings of the AAAI/ACM Conference on AI, Ethics*, and Society, 2019, p. 195-200, *https://doi.org/10.1145/3306618.3314289; Scarpino; An Exploratory Study*

36 Casey; Farhangi; "Rethinking Explainable Machines"

37 Muller, H.; Mayrhofer, M. T.; et al.; "The Ten Commandments of Ethical Medical AI," *Computer*, vol. 54, iss. 7, 2021, p. 119–123, *https://doi.org/10.1109/MC.2021.3074263*