Evaluating Ethical Challenges in AI and ML

Disponible également en françaiss
www.isaca.org/currentissue

In general, privacy, bias and discrimination are currently receiving a lot of attention. However, it is common for them to be underprioritized in technology implementations and treated as isolated issues, only receiving attention when necessary. Many organizations instead prioritize goals such as efficiency gains or increased profits, which often require richer data sets, but they fail to consider the eventual impact of their data handling methodologies on foundational social justice issues.[1] The consequences of implementing technologies without fully understanding the privacy, bias and possible discrimination issues they pose threaten both individuals and enterprises. Built-in bias can negatively affect an individual's ability to receive fair treatment in society. For organizations, the negative potential includes reputational damage, financial impact, litigation, regulatory backlash, privacy concerns and diminished trust from clients and employees.[2] Developers of technology applications should aim for them to be impartial, unbiased and neutral, and organizations should consider these foundational issues during the implementation of emerging technologies to ensure that bias and discrimination are not fundamental components of a system's design.

Ethical Behavior

Ethics are generally defined as a set of standards for determining what behavior is considered right and wrong in a particular group, culture or society based on accepted norms.[3, 4] Although there is often consensus regarding some behaviors—lying and cheating are typically considered unethical—opinions about what constitutes ethical behavior can sometimes diverge dramatically from one culture to another. Some of the ethical dilemmas that relate to technology include using artificial intelligence (AI) to replace humans in performing certain roles and leveraging these systems to make automated decisions within organizations with little to no oversight, ultimately resulting in possible adverse outcomes for society.

Systemic Issues

With respect to implementations of AI and machine learning (ML), ethics issues have come to a critical inflection point, requiring organizations to balance operational goals with individual rights.[5] Trust is a component of ensuring confidence in technology; knowing that a system made a decision at the right time for the correct reason is critical. With this comes the foundational need for a system to be explainable so that it is easy to articulate why the system made a given decision and maintain a high level of confidence in the design.[6, 7]

One article proposed 10 practical guidelines for the application of AI to a broad group of stakeholders. While the focus was on the use of AI in medical cases, the guidelines lend themselves to universal application. Among other things, they aim to ensure that technological operations can be easily explained, designs are transparent, decisions are recognizable and repeatable, and humans take ownership of these decisions.[8]

Two of the ten guidelines address relevant foundational issues:

JOSHUA SCARPINO | CISM, CISSP

Is the global manager of security and compliance at Harver, where he leads the global security and compliance team. He has more than 18 years of experience in IT and security and 16 years of experience in the US Air Force. He has overseen security operations for Fortune 500 enterprises and enhanced critical controls at financial organizations deploying, establishing, scaling and auditing their security programs to attain compliance internationally. Throughout his career, Scarpino has bridged multiple security areas to resolve operational, governance, risk and compliance challenges. He has a passion for education and currently teaches part time as an adjunct instructor and is conducting research for his doctoral program on ethical considerations for artificial intelligence and machine learning at Marymount University (Arlington, VA, USA).

1. "An AI decision, action, or communication must not violate any applicable law and must not lead to human harm."
2. "An AI decision, action or communication shall not be discriminatory. This applies in particular to the training of algorithms."[9]

Despite many efforts to identify what is needed to comply, with some contributors even providing frameworks to guide deployments, there is still no reliable method to identify and aid in prioritization when the risk of harm, discrimination or other ethical concerns may exist. And although consequences are significant for AI and ML when deployed at scale, this potential risk is not limited to these technologies.

Though there is increased awareness, there is not yet a unified approach for identifying these systemic issues, and there are no standardized procedures to address them consistently.

In recent years, individual privacy rights have drawn increased attention, and social justice awareness has become a core discussion point in many regions. At the same time, the need to adopt universal standards to ensure ethical technology implementations has grown. As AI and ML capabilities evolve, a standardized approach is needed to determine if an organization is subject to additional risk. It is imperative that enterprise leaders understand the consequences of failure to mitigate the risk associated with a lack of controls to address privacy, bias and discrimination challenges in AI and ML technologies. Though there is increased awareness, there is not yet a unified approach for identifying these systemic issues, and there are no standardized procedures to address them consistently. This is a foundational problem for many organizations around the world. Leaders must understand the importance of these issues and act appropriately to eliminate bias and discrimination from all technology implementations.[10]

Creating Trusted Systems

Current industry examples and expert opinions can be used to determine whether organizations would benefit from quantifying the potential privacy, bias and discrimination risk present in their technology implementations. A model can be used to understand how organizations currently handle this risk and highlight the benefits of a unified approach that implements technological controls while raising awareness of potential ethical concerns associated with these fundamental social justice issues. The benefits of developing such a model and assisting organizations in highlighting these critical issues during technology implementation and design phases are evident, compared to the disadvantages of implementing technologies without considering these fundamental issues. It is essential to examine the relevant requirements for a given system and current approaches that an organization may be leveraging to understand those risk areas prior to implementation to ensure that critical issues regarding potential bias are appropriately addressed. A preliminary review can provide a starting point for a conversation, help increase awareness of these problems and provide a basis for remediation efforts that will mitigate the identified risk.[11]

A review of some of the current literature regarding how organizations currently approach implementation with respect to foundational issues reveals significant concerns.[12] One example describes an Amazon implementation of AI that used a hiring algorithm that was found to be biased. The system favored words such as "captured" or "executed." These words were more often found on male resumes, which led to the algorithm favoring male applicants. This AI deployment unfairly limited participation by female job applicants. Although

Amazon has corrected the problem, the individuals affected were likely not offered remediation for harm caused.[13] This example highlights just one of the many ways deployed technologies can contribute to foundational problems. Allowing decision-making based on built-in bias not only causes harm to those directly impacted but can also increase distrust in these systems. Another example is a widespread view that discrimination is built into the evaluation system Classic Fair Isaac Corporation (FICO) model.[14] Critics contend it favors white Americans over people of color because it values traditional credit more than positive payment records. Aracely Panameño, director of Latino affairs for the Center for Responsible Lending, noted that "If the data that you're putting in is based on historical discrimination, then you're basically cementing the discrimination at the other end."[15]

Many ML algorithms are difficult to explain and derive how the answer was obtained by the system.

Although organizations claim data are impartial, they are often unable or unwilling to provide proof of their claims.[16] Many ML algorithms are difficult to explain and derive how the answer was obtained by the system; these are known as black box systems. The results are based on assumptions for how the systems come to a decision.[17, 18] These systems have far-reaching impacts. Research studying a large portion of publications shows that many topics they discuss only partially relate to "explainable, accountable and intelligible systems."[19] The only categories identified to have any relationship to ethics and privacy are "big data privacy, trust, algorithmic fairness, and explanation and reasoning." These focus areas represent a small proportion of all research completed in this field.[20] There is a significant lack of focus on ethical and discrimination-based research in AI. When AI and related systems are designed and implemented, it is crucial to understand how they might contribute to decision-making that could have ethical implications within an organization.

The Path Forward: A Consistent Approach

As technology evolves, preventing bias, promoting data privacy and protecting inherent human rights are paramount. Critical to conversations about these foundational issues is the recognition that humans are inherently biased, and there is a risk that anything human created will have bias as a component.[21] Despite efforts to design AI and ML systems that are free from or minimize bias, it may, nevertheless, exist, either due to personal beliefs held by the creators, cultural biases or built-in discrimination in the way data sets are used to train systems. Organizations must make an effort to fundamentally understand how the systems they design could impact users.[22]

In researching AI/ML development life cycles, most publicly shared development life cycles share or resemble the steps depicted on the left of **figure 1**. One distinct component to combat bias in federal systems is the completion of an impact assessment for bias prior to development and implementation, determining the appropriateness of the solution, and validating iteratively throughout the deployment process.[23] Fundamental changes in how these systems are created, designed and implemented are needed. The addition of distinct steps to address bias during the AI/ML development life cycle would benefit all organizations and promote a consistent approach to identifying these fundamental issues. Adding two distinct steps—one to assess bias potential and resulting ethical implications and one to validate risk after deployment—could go a long way toward ensuring ethical challenges are considered during these technology implementations. The proposed additions are depicted on the right in **figure 1**.

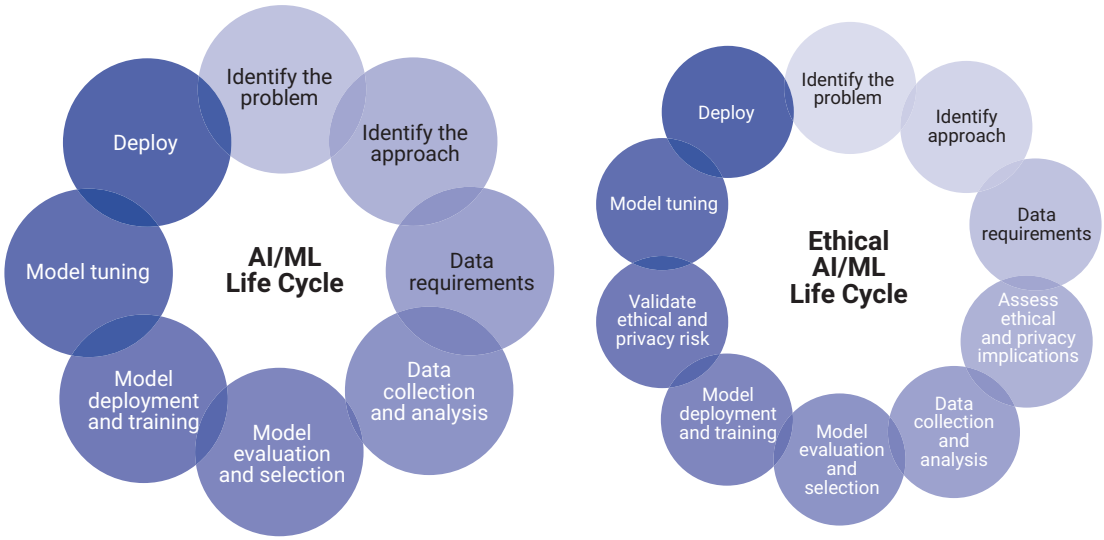
Assessing ethical and privacy implications requires organizations to consider and document these concerns as part of the development process. Although there are numerous ethical AI frameworks to guide the development and evaluation of ethical AI, these distinct steps have not been universally adopted into the general AI development life cycle, and they are typically an addition to the process. Making this a core component of the life cycle for all AI and ML implementations is critical to universal adoption and success. The validation of ethical and privacy risk is necessary to ensure that the

LOOKING FOR MORE?

- Read *AI Uses by Blue Teams Security. www.isaca.org/ai-blue-team-security*
- Learn more about, discuss and collaborate on emerging technology in ISACA's Online Forums. *https://engage.isaca.org/onlineforums*

FIGURE 1

Typical vs. Ethical AI/ML Life Cycle



considered and evaluated risk has not changed once a system has been developed and deployed.

The challenge is not that the tools and frameworks are unavailable or that individuals do not care about these issues. An online search for "ethical AI" frameworks reveals a significant number of resources and experts that promote ethical approaches. The current challenge is that many organizations remain focused on meeting business or operational objectives and ensuring project deadlines and budgets remain on target. It is likely that concerns about bias are not intentionally discarded—they simply are not adequately prioritized due to operational goals.

Conclusion

For this conversation to advance, current industry behavior patterns need to be addressed. Favoring operational goals and efficiencies without prioritizing ethics is no longer an acceptable approach. The climate may be changing, as revealed by the challenges some large enterprises have faced over the biases designed into their systems and deployed at scale. Some have reevaluated and modified their technology implementations to address built-in bias and refined their approaches. Designers of systems must acknowledge the tendency to impart their individual biases into systems. Further, enterprises must realize that when technology implementations have built-in biases, using them to leverage data sets may perpetuate discrimination at scale. Only by understanding where these fundamental problems arise will it be possible to develop nondiscriminatory systems and mitigate risk.

One promising path forward is to adopt and add distinct steps to the AI and ML life cycle, ensuring that ethical and privacy concerns are fundamental components of all design and development processes for AI and ML. Furthermore, a process that validates the potential residual risk after development but before implementation to ensure expected outcomes is vital. For sustainable confidence in technology and organizations, it is vital to ensure that emerging AI and ML systems uphold individual rights to fairly and equitably participate in society.

Endnotes

1 Lo Piano, S.; "Ethical Principles in Machine Learning and Artificial Intelligence: Cases From the Field and Possible Ways Forward," *Humanities and Social Sciences Communications*, vol. 7, iss. 1, 17 June 2020, *https://www.nature.com/articles/s41599-020-0501-9*

2 Cheatham, B.; K. Javanmardian; H. Samandari; "Confronting the Risks of Artificial Intelligence," *McKinsey Quarterly*, 26 April 2019, *https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/confronting-the-risks-of-artificial-intelligence*

3 Singer, P.; "Ethics Philosophy," Britannica, 15 December 2021, *https://www.britannica.com/topic/ethics-philosophy*

4 IGI Global, "What Is Ethics," *https://www.igi-global.com/dictionary/ethics-in-higher-education/10276*
5 *Op cit* Lo Piano
6 *Ibid*.
7 Muller, H.; M. Mayrhofer; E. Van Veen; A. Holzinger; "The Ten Commandments of Ethical Medical AI," *Computer*, July 2021, *https://ieeexplore.ieee.org/document/9473208*
8 *Ibid.*
9 *Ibid*.
10 Liu, X.; D. Murphy; "A Multi-Faceted Approach for Trustworthy AI in Cybersecurity," *Journal of Strategic Innovation and Sustainability*, vol. 15, iss. 6, 16 December 2020
11 Munoko, I.; H. L. Brown-Liburd; M. Vasarhelyi; "The Ethical Implications of Using Artificial Intelligence in Auditing," *Journal of Business Ethics*, vol. 167, iss. 2, 8 January 2020, *https://link.springer.com/article/10.1007/s10551-019-04407-1*
12 Trunk, A.; H. Birkel; E. Hartmann; "On the Current State of Combining Human and Artificial Intelligence for Strategic Organizational Decision Making," *Business Research*, 20 November 2020, *https://link.springer.com/article/10.1007/s40685-020-00133-x*
13 Manyika, J.; J. Silberg; B. Presten; "What Do We Do About the Biases in AI?" *Harvard Business Review*, 25 October 2019, *https://hbr.org/2019/10/what-do-we-do-about-the-biases-in-ai*
14 Consumer Financial Protection Bureau, "What Is a FICO Score?" 4 September 2020, *https://www.consumerfinance.gov/ask-cfpb/what-is-a-fico-score-en-1883/*
15 Martinez, E.; L. Kirchner; "The Secret Bias Hidden in Mortgage-Approval Algorithms," The Markup, 25 August 2021, *https://themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms*
16 *Op cit* Manyika *et al.*
17 *Op cit* Liu
18 *Op cit* Lo Piano
19 Abdul, A.; J. Vermeulen; D. Wang; B. Lim; "Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda," Proceedings of the 2018 Conference on Human Factors in Computing Systems, 21 April 2018, *https://dl.acm.org/doi/10.1145/3173574.3174156*
20 *Ibid*.
21 Livingston, M.; "Preventing Racial Bias in Federal AI," *Journal of Science Policy and Governance*, vol. 16, iss. 2, May 2020
22 Wallach, W.; "Robot Minds and Human Ethics: The Need for a Comprehensive Model of Moral Decision Making," *Ethics and Information Technology*, vol. 12, iss. 3, September 2010
23 *Op cit* Livingston