ACTIONABLE ETHICS FOR FAIRNESS IN AI

By Cathy Petrozzino and Stuart Shapiro

Artificial intelligence (AI) is an increasingly pervasive tool with unintended consequences. AI technology and the tools that rely on it have been growing in use in all facets of society.

From consumer uses such as whole-home automation systems to applications in manufacturing, space, and national security, AI is contributing to operational efficiencies. However, there is also growing awareness of AI solutions that have resulted in disparate impacts and harm. As AI becomes more ubiquitous, concerns about how and when it is used take on greater urgency and importance.

These concerns should not sideline AI tools, which offer potential benefits. Rather, these concerns should be acknowledged and addressed by actively working to build fairness and ethical logic into AI solutions and by stress-testing these solutions for potential unintended consequences. This requires organizations to institutionalize "proactive ethics" into the AI lifecycle, which includes analyzing the ethical implications for stakeholders at all levels: individuals, groups, society, and the organization and its employees.

Moving forward with ethical uncertainty poses harm to all stakeholders. Addressing the AI ethical challenge now is critical for establishing AI as a vehicle for creating a safer, more efficient, and more equitable world.

Ethical Debt and Its Consequences

Ethical debt is incurred when an agency opts to design, develop, deploy, and use an AI solution without proactively identifying and mitigating potential ethical concerns. Unlike technical debt, the burden of ethical debt is paid only in part by the agency, but often to a larger degree by stakeholders outside the agency. Ethical debt further accumulates when AI models or results are reused or repurposed based on an erroneous assumption that the model addresses ethical concerns, including fairness, consistent with the agency's ethics priorities.

In recent years, private- and public-sector organizations strove to derive value from the vast quantities of available public and private data. During this period of AI innovation, disturbing observations have been made, particularly in terms of fairness and bias.

- Independent analysis demonstrated that a popular commercial algorithm used to identify patients who require extra healthcare support favored healthier White patients over sicker Black patients.
- Many AI-based hiring models that were designed to combat discriminatory hiring "drifted" over time to reinforce existing biases against hiring women and other underrepresented groups.
- Due to "automation bias," AI tools designed to assist decision-makers became the decision-makers, reducing the efficacy of the "human-in-the-loop" control.

MITRE, a not-for-profit company that operates federally funded R&D centers, has worked with AI for decades,

more recently in the form of sophisticated data analytics and machine learning. MITRE has conducted broad and sector-specific research on AI ethics with a detailed focus on fairness. MITRE recognizes that for an AI solution to serve the public good, it needs to "think" beyond technical and functional features, and also address consequences that may create inequities, harm, and other ethical issues. MITRE has used its expertise and independence to help sponsors incorporate ethics into their AI solutions. This includes enhancing the existing AI lifecycle via the Lifecycle Ethical Analysis (LEA) methodology to achieve actionable ethics.

Operationalizing Ethics within AI in Federal Agencies

Principles and Guidance are Essential but Not Enough

To date, solutions to the AI ethics challenge have focused on high-level principles and frameworks, or detailed adjustments to AI models. The Department of Defense's (DoD) Joint Artificial Intelligence Commission has established five ethical principles to underpin the DoD's AI efforts. Internationally, the United States helped lead the Organisation for Economic Cooperation and Development's effort to define AI principles.

A number of organizations have gone beyond principles to more detailed guidelines and frameworks. Google and Microsoft have implemented governance structures to support ethics in AI. The Institute of Electrical and Electronics Engineers (IEEE) has published a collaborative exploration of Ethically Oriented Design for AI systems.

There has also been considerable effort, particularly in the academic community, to address fairness by adjusting models to compensate for known statistical biases. With these adjustments, an AI solution can be designed to counter human and data biases and help produce a more equitable result.

Principles and guidance are essential stepping-stones to accomplishing ethics, but they lack the detail to support AI developers in identifying and evaluating ethical issues. Statistical adjustments can help improve model results, but may miss the context of how the model is used and the potential impact of the adjustments in practice.

High-level principles, frameworks, and detailed adjustments to models are more effective in addressing ethical debt when coupled with proactive leadership and engineering, plus an effective LEA process.

Proactive Leadership Response

Agency leaders and policymakers who are aware of the limitations associated with AI are able to establish an agency that is empowered and committed to developing a culture of AI solutions that are demonstrably ethically grounded.

Proactive leadership is necessary so that critical policy considerations are addressed, beginning with the ethical principles (i.e., values), which will drive the agency's AI ethics program. After the principles have been identified, they need to be formalized across the AI ethics program from the overarching policies to the standards used in the AI lifecycle phases. The policies must be broadly positioned to consider external impact to all stakeholders and associated equities, including those of individuals, families, vulnerable groups, society, and other organizations.

Policies and training must establish and enforce a culture of ethics. Ethical issues in an AI solution may first be noticed by a technical team member with little agency clout. Policies must allow team members to feel safe when airing concerns and to know that their concerns will be addressed.

Given the potential impact of AI solutions, organizations need to incorporate healthy, ethical questioning during the conceptualization of an AI-based solution. The first imperative for any kind of actionable ethics is to ask whether a particular AI-based system or application should be pursued at all. The "what ifs" should be examined in the context of its prospective use by a diverse team, both demographically and in terms of skills. The importance of such existential questioning has long been recognized, but it is often neglected in practice.

For leaders who are responsible for AI solutions, particularly consequential AI, the stakes are high. But deploying an AI-based system that is inequitable

and ethically indefensible serves neither the mission nor the organization and can cause real harm. Knowledge, effective questioning, and establishing a culture of AI ethics through principles and policies enables leaders and policymakers to effectively navigate ethics in the highly technical AI space.

Proactive Engineering Response

Application of ethical reasoning in socio-technical contexts currently falls into two principal modes. One is the discipline of engineering ethics, which serves as both a basis for postmortems of socio-technical controversies or disasters and as a proactive preventative. The other is impact assessments.

The importance of engineering ethics is acknowledged in the form of educational accreditation requirements (e.g., the Accreditation Board for Engineering and Technology [ABET]) and professional society ethical codes (e.g., the Association for Computing Machinery [ACM]). However, effectively equipping systems engineers, computing professionals, and policymakers with the skills to apply explicitly ethical reasoning continues to be a work in progress.

Impact assessments, whether or not explicitly designated as ethical analyses, frequently incorporate some sort of ethical reasoning, even if only implicitly. Some of these focus on specific aspects of socio-technical systems, such as algorithmic accountability or privacy. Others take a broader social impact and/or human rights perspective.

However, more is needed to support agencies' efforts to address ethics in AI solutions in an actionable way that factors in essential contextual considerations. Further, the complexity of these solutions demands an approach beyond (but consistent with) the traditional modes of engineering ethics and impact assessment.

Lifecycle Ethical Analysis

Integrating ethical reasoning with systems engineering and computer science AI practices should not be interpreted as rendering ethics as simply another lifecycle activity. Ethical reasoning is dynamic and contextual and must occur at every stage of the AI development lifecycle, not as a component of it, but concurrent with it. The Lifecycle Ethical Analysis (LEA) process illustrated below provides a basis for ethical reasoning alongside AI system development.

Ethical reasoning has to be translatable into technical frames and vice versa. Translations between ethics and technology need to be more effectively supported and understood. There are multiple conceptual links between ethics and AI technology under an overarching concept of trustworthiness, including safety, security, privacy, and fairness. There will, in effect, be an ethics lifecycle and a systems lifecycle running concurrently to inform each other. These two lifecycles must be connected by a feedback loop to assess whether the system is working within acceptable tolerances, to monitor for unintended

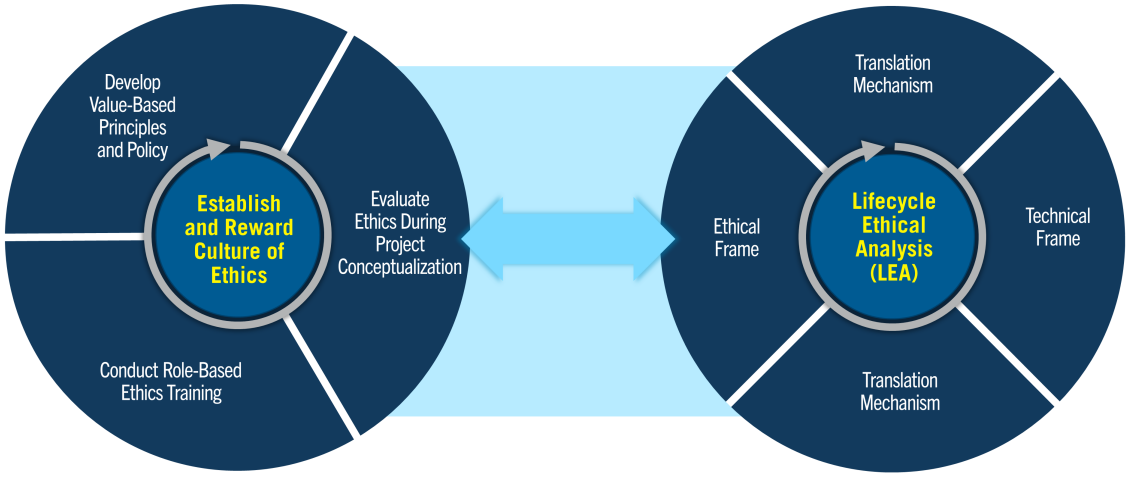

Figure 1. Addressing the AI Ethics Challenge

consequences, and to validate that system functions reflect the agency's ethical principles.

LEA would be supported by appropriate tools and methods and tailored by lifecycle stage for optimal effectiveness. Existing methodologies such as counterfactual analysis and assurance cases offer the potential to act as translators between ethical concepts and system properties. Other methods may need to be developed to serve as the means by which ethical reasoning comes about; for example, fairness can influence functional reasoning about fairness and vice versa.

Recommendations

Agency Actions for Achieving Ethics in AI

Before agencies can run, they need to walk. Key prerequisites that largely fall under the auspices of leadership and policymakers include:

- Ethical principles must be created that drive organizational policies and procedures that support ethics analysis and open dialogue.
- Role-based AI ethics training and awareness should be mandated for agency leaders, policymakers, developers, and users of AI-based solutions.
- Diverse and multidisciplinary AI teams should be established to analyze ethics from a broad stakeholder perspective.

These actions will establish a culture of ethics and the ground upon which agencies can begin to run rather than walk. Once established, the next steps are for agencies to develop and institutionalize LEA, aided by overarching guidance from the Office of Management and Budget, by:

- Adjusting existing impact-assessment vehicles (e.g., privacy impact assessments) to reflect the agency's AI ethical principles.
- Identifying and adapting relevant framing methodologies to support translation between ethical and technical concepts.
- aligning LEA with existing AI lifecycle activities in specific environments.

Beyond individual agencies, a cross-agency coalition of AI developers and users that are committed to ethics in AI should:

- Share approaches that have led to demonstrable fairness and trust in AI solutions.
- Collaborate on research to develop methods and tools for supporting LEA.
- Provide an independent knowledgeable perspective to help analyze ethics in challenging AI solutions and contexts.

Conclusion

The missions of agencies align with solid ethical principles as opposed to ethical debt. What agency leader would support an AI solution that systemically makes biased predictions, or is more harmful than beneficial? What policymaker would endorse the deployment of an AI solution that conflicts with the letter and the spirit of the organization's principles? Complex socio-technical systems, however, require more than good intentions. They require frameworks and methods that effectively enable the proactive application of ethical reasoning to agency operations involving AI. This is doubly imperative, as ethics is not just a matter of doing good; it's tightly coupled with good mission outcomes and creating an efficient, safer, and more equitable world for all.

For more information about this paper or the Center for Data-Driven Policy, contact us at policy@mitre.org