AI FORUM

Year: 2025 | Edition: 17 |
Cambodia | 16 May 2025 |

INSIGHTS

Theravada Buddhism: Mindfulness for the Ethical Advancement of AI

CHEAB Sambath[1], MD

CHHEM Kieth Rethy[2], MD, PhD (Edu), and PhD (His)

[1] ***CHEAB Sambath*** *is a Lecturer at University of Puthisastra, Cambodia*
[2] ***CHHEM Kieth Rethy*** *is a Distinguished Professor of AI and Humanities, Center for Buddhist AI Ethics, Cambodia University of Science and Technology and Advisor to AI Forum*

Abstract

Artificial intelligence (AI) raises significant ethical concerns regarding justice, transparency, and responsibility, demanding novel approaches to mitigate risks like algorithmic discrimination and social exclusion. This commentary investigates how Theravada Buddhist mindfulness (sati) offers a transformative ethical foundation for AI development. Mindfulness fosters self-awareness and wisdom, enabling developers to identify biases, such as those in hiring systems reinforcing inequality, and design systems promoting equity and compassion. By bridging Buddhist ethical principles with insights from modern AI ethics scholarship, this approach provides a valuable framework for developing technology that minimizes harm and uplifts humanity.

Overview of Theravada Buddhism and Ethical Artificial Intelligence

Theravāda Buddhism, the "Teaching of the Elders," offers a clear ethical framework centered on sīla (ethical conduct), samādhi (mental cultivation), and paññā (wisdom), guiding individuals to overcome suffering (dukkha) caused by greed (lobha), hatred (dosa), and delusion (moha). Liberation (Nibbāna) arises through insight into impermanence (anicca), unsatisfactoriness (dukkha), and non-self (anattā) (Gethin, 1998). This framework speaks directly to ethical concerns in AI. Unchecked AI development, often driven by ego, self-interest, or ignorance, mirrors unwholesome mental roots, leading to bias, exclusion, and harm—e.g., discriminatory facial recognition or surveillance misuse. The Buddhist call for mindfulness and ethical restraint offers a corrective lens. Recent AI ethics literature echoes this. Bostrom (2014) highlights risks of misaligned superintelligence; Floridi (2015) frames AI within a broader moral infosphere; and Whittlestone et al. (2019) advocate integrating ethics throughout AI design, not as a post-hoc fix. Theravāda principles align with these views, particularly through their shared emphasis on intentionality, responsibility, and long-term well-being. Integrating mindfulness into AI development fosters humility and ethical clarity, complementing global efforts to create AI that serves rather than exploits. Theravāda ethics thus provides both a cautionary and constructive perspective within the broader landscape of responsible AI.

Defining Mindfulness and Its Significance in AI Development

In Theravāda Buddhism, sati (mindfulness) is the non-reactive, lucid awareness of phenomena as they unfold in the present moment, free from distortion, bias, or self-referential conceptualization. Grounded in the Satipaṭṭhāna Sutta, it involves direct observation of the body, feelings, mental states, and mental objects, enabling one to perceive things as they truly are (yathābhūtañāṇadassana)—impermanent, conditioned, and not-self. For example, in ethical AI design, a developer may believe they are creating a "neutral" algorithm. However, through mindful awareness of their own assumptions, intentions, and social conditioning, they may come to see how bias—such as prioritizing profit over fairness or overlooking marginalized users—actually shapes the system. Sati thus helps distinguish between the delusion of neutrality and the reality of embedded ethical consequences.

This practice helps practitioners recognize and avoid unwholesome mental states (greed, hatred, delusion) and understand reality's interconnected, impermanent nature. While this article focuses on Theravada *sati*, it's worth noting that mindfulness practices are also central to other Buddhist traditions, such as Mahāyāna, and have been adapted into secular contexts, sometimes with different applications or goals. As AI evolves, mindfulness offers a crucial tool for developers to maintain ethical integrity and self-awareness. A mindful developer consciously examines the origins and potential societal impacts of their work. This involves cultivating the awareness to question algorithmic fairness and inclusivity. Furthermore, mindfulness practices can enhance well-being, reduce stress, and foster the resilience needed to navigate the complexities of ethical AI creation with compassion and wisdom.

Mitigating Bias with Mindfulness: Understanding the Four Biases (Agati)

Algorithmic bias is a pervasive challenge in AI. Theravada Buddhism offers a framework for understanding bias through the concept of *agati*, or prejudice, identifying four key types: bias arising from desire or partiality (chanda), aversion or hatred (dosa), fear (bhaya), and delusion or ignorance (moha) (Thanissaro Bhikkhu, 2002). Algorithmic bias is a pervasive challenge in AI. Theravāda Buddhism offers a powerful ethical lens through the concept of *agati*—deviation from just or righteous action—highlighting four key types of bias:

- **Chanda-agati (bias from desire or attachment):** When a developer or organization favors outcomes due to personal gain, loyalty, or profit, objectivity is compromised—e.g., designing algorithms to optimize engagement even if they spread misinformation.
- **Dosa-agati (bias from aversion or hatred):** This occurs when decisions are driven by resentment, competition, or implicit hostility—such as excluding certain groups from datasets or dismissing dissenting ethical critiques.
- **Bhaya-agati (bias from fear):** Fear of failure, reputational damage, or regulatory scrutiny can lead to cutting corners or avoiding ethical transparency in AI deployment.
- **Moha-agati (bias from delusion or ignorance):** Perhaps the most pervasive, this bias stems from a lack of awareness—e.g., ignorance of social inequalities encoded in data, or blind faith in "neutral" algorithms.

These distortions can unconsciously shape AI systems. Mindfulness enables developers to observe these mental states and motivations impartially as they arise. This cultivated awareness fosters the clarity needed to identify and proactively address potential biases in datasets, algorithms, and deployment strategies, thereby promoting the design of fairer and more just AI systems.

Practical Implications and Addressing Challenges

While originating in Buddhism, the core principles of mindfulness – awareness, compassion, nonjudgmental observation – have universal applicability. However, integrating these principles into the fast-paced, often secular, tech industry presents challenges. Critics may question the feasibility or raise concerns about cultural appropriateness or potential superficiality ("ethics washing"). Effective integration requires more than just isolated workshops; it necessitates fostering an ethical culture. Practical approaches include structured meditation sessions, ethical reflection seminars, and mindfulness-oriented training designed to enhance self-awareness and ethical decision-making. These methods help developers consider the broader societal repercussions of their work. Furthermore, integrating mindfulness principles conceptually into existing AI ethics guidelines (like IEEE standards or the EU AI Act) can ensure ethical considerations remain central. Companies like Google have explored mindfulness programs, suggesting potential benefits, but challenges remain in ensuring genuine integration rather than superficial adoption. Scaling these practices effectively across diverse teams and contexts requires careful consideration and commitment from organizational

leadership. The goal is to create environments where ethical reflection and societal impact are intrinsic parts of the development lifecycle.

Ethical and Societal Implications of Mindful AI Creators

A mindful AI creator approaches their work with humility, wisdom, and empathy, recognizing its impact within an interconnected world. They prioritize long-term ethical considerations and societal benefit over personal gain or unexamined biases. The prevalence of ethical issues in AI, such as biased facial recognition or discriminatory algorithms, underscores the critical need for such responsible development. Mindfulness cultivates the ethical sensitivity required to actively address these problems. A mindful developer would critically assess how a technology might impact vulnerable groups and build in safeguards. This approach ensures AI systems are developed with fairness, transparency, and accountability, upholding human dignity and aligning with robust ethical standards. Developers thus become ethical stewards, actively working to ensure technology serves humanity.

Conclusion: Call for Action

Integrating mindfulness (sati) with AI development offers a powerful ethical framework for navigating contemporary technological challenges. By fostering self-awareness, compassion, and wisdom, mindfulness empowers developers to recognize and mitigate biases and unwholesome motivations, ensuring AI aligns with justice, transparency, and societal benefit. While rooted in Theravada Buddhism, its core tenets are broadly applicable and can complement existing ethical frameworks and training programs. As AI continues to shape our world, developers, educators, policymakers, and organizations must consider embracing mindfulness not just as a technique, but as a fundamental principle for ethical AI. Cultivating awareness, compassion, and responsibility is essential to ensuring AI serves equity, harmony, and human flourishing. Let us commit to building a future where technology uplifts, rather than harms, humanity.

References

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

Floridi, L. (2015). *The onlife manifesto: Being human in a hyperconnected era*. Springer. https://doi.org/10.1007/978-3-319-04093-6

Gethin, R. (1998). *The foundations of Buddhism*. Oxford University Press. https://global.oup.com/academic/product/the-foundations-of-buddhism-9780192892232?cc=us&lang=en

IEEE. (2025). IEEE standard for algorithmic bias considerations (IEEE Std 7003-2024). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/IEEESTD.2025.10851955

Thanissaro Bhikkhu. (2002). *The four biases* [PDF]. Dhammatalks.org. https://beta.dhammatalks.org/Archive/Writings/Transcripts/0210n1a2FourBiases,The.pdf

Whittlestone, J., Nyrup, R., Alexandrova, A., & Cave, S. (2019). The role and limits of principles in AI ethics: Towards a focus on tensions. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 195–200. https://doi.org/10.1145/3306618.3314289

** Edited with ChatGPT for clarity; all views expressed are solely those of the authors.*

** The opinions expressed are the authors' own and do not reflect the views of AI Forum.*