Chapter : 5

LOGIC OVER RHETORIC: A NYAYA INFERENTIAL ANALYSIS OF AI ETHICS

Subhodeep Mukhopadhyay [1]

ISBN: 978-91-410022-2-7 | **DOI:** 10.25215/9141002229.05

Abstract:

The artificial intelligence (AI) industry has witnessed a rapid proliferation of AI ethics frameworks, in line with the widespread global adoption of AI. Such guidelines are meant to act as safeguards and typically seek to address issues like accountability, fairness, transparency, and justice. However, their practical effectiveness has been a subject of intense debate. This paper examines three such structural limitations of contemporary AI ethics frameworks: lack of enforceability, functional vagueness, and corporate capture. Using the *Nyaya pancha-avayava* (five-membered) inferential framework as an analytical lens, the study reformulates these critiques into structured logical propositions. The purpose is to find out if the existence of a code of conduct can realistically prevent harm. The analysis demonstrates that, in the absence of *udaharana* or verifiable empirical examples of non-binding functionally vague corporate-driven principles successfully restraining profit-driven harm, there can be no behavioral change. The findings provide a formal logical foundation of ethics-washing and show that until ethics frameworks satisfy the rigors of formal inference, they remain empty epistemic exercises rather than functional tools of justice.

Keywords*: AI Ethics, Nyaya, Pancha-Avayava, AI Governance, Artificial Intelligence*

[1] Senior Research Fellow, Infinity Foundation India, Chennai, India, Email Id: subhodeepm.infinity@gmail.com

Introduction:

Today Artificial Intelligence (AI) is used across different domains such as governance, healthcare, finance, education, and media. The scale at which AI systems influence decisions is massive, often cross-cultural and trans-national. Consequently, ethical concerns related to accountability, transparency, fairness, responsibility, and harm have gained prominence (IBM, 2021; ISO, 2025). Other major socio-economic and policy concerns include the future of jobs and environmental impact of AI (Mukhopadhyay, 2025). The ethics of AI has thus emerged as a rapidly growing field of research. Recent scholarship has attempted to map, classify and identify these concerns systematically. Huang et al. (2023), for example, lists sixteen ethical issues and organizes them across three dimensions: individual, societal, and environmental.

A large number of AI ethics frameworks have been issued by technology companies and professional bodies. However, concerns have been raised over their practical effectiveness (Corrêa et al., 2022). There are a number of reasons for this. It is often difficult to translate high-level ideas into concrete practices. There are also limited mechanisms for independent oversight. Moreover, ethics frameworks are often designed after the fact, and do not necessarily not drive the AI systems development process. Scholars have also pointed to other persistent problems. These include lack of enforceability (Khan et al., 2022) and the fact that frameworks are often vague (Hagendorff, 2020). Scholarship also suggests that AI codes of conduct, more often than not, serve corporate interests (Fioravante, 2024). This makes them ineffective in addressing systemic and structural harms. Against this background, the present paper analyzes contemporary AI ethics frameworks using the *Nyaya pancha-avayava* inferential structure. By reformulating key critiques in this classical logical framework, the paper demonstrates why many existing AI ethics guidelines are structurally incapable of preventing or stopping harm, despite their normative aspirations.

Methodology:

This study employs the *Nyaya pancha-avayava* framework. *Nyaya* is one of the six schools of Indian philosophy and deals mainly with logic and reasoning. It provides a systematic framework for analysis and debate. In order to do this, *Nyaya* has developed a rigorous framework to define what valid knowledge (*prama*) is and the means by which it

can be gained (*pramana*). For this reason, it is also known as *pramana-shatsra* or the science of reasoning. The main idea behind the *Nyaya* inferential method is that conclusions must be based on clear logic, and not on assertions or intuition (Gopinath & Sharma, 2022).

Pancha-avayava is a five-step inferential structure. It consists of proposition (*pratijna*), reason (*hetu*), general rule with example (*udaharana*), application (*upanaya*), and conclusion (*nigamana*). This structure helps establish valid inference through a logical relationship between a reason and a conclusion (Sarukkai, 2005). To illustrate the framework, the standard example of fire and smoke is seen in the *shastras*. The hill is inferred to have fire because it has smoke (*pratijna* and *hetu*); wherever there is smoke, there is fire, as in a kitchen (*udaharana*); the hill has smoke of that kind (*upanaya*); therefore, the hill has fire (*nigamana*).

Building on this structure, the study applies the *pancha-avayava* framework to analyze some of the current criticism of AI ethics frameworks. Each claim is reformulated as a *nyaya* style inference, and studied to see if there are well-accepted parallel examples (*udaharana*) which justify the basic premise.

Results:

The results presented next are from a formal *pancha-avayava* analysis of the three AI ethics critiques alluded to earlier: lack of enforceability, functional vagueness, and corporate alignment. This allows one to logically establish that such critiques are justified and that AI guidelines based on them cannot prevent harm.

(a) Lack of Enforceability:

Here we examine the ineffectiveness of non-binding AI ethic principles to prevent harm. In this inference, the *hetu* (reason) for failure is the absence of enforceability, which is a prerequisite for restraining profit-driven behavior.

- **Pratijna**: AI ethics frameworks cannot prevent systemic harm.
- **Hetu**: Because they are non-binding (un-enforceable) in nature.
- **Udaharana**: Whatever is non-binding, cannot prevent harm. Just as suggestions without enforcement do not deter crimes.

- **Upanaya:** And AI ethics frameworks are not enforceable.
- **Nigamana:** Therefore, AI ethics frameworks cannot stop harm.

This shows that non-enforceability of AI ethics guidelines makes them ineffective in actually preventing harm. The logic is validated through the universal rule (*udaharana*) that a suggestion without the power of sanction (*danda)* cannot deter someone from carrying out a criminal act. This is noted in the Artha-shastra (1.4.13-14) that without the law of punishment (*danda*), the law of the fish (*matsya-nyaya*) prevails and the strong swallows the weak:

"apraṇītaḥtumātsyanyāyaṃudbhāvayati |

balīyānabalaṃ hi grasatedaṇḍadharābhāve |"

This leads to the inevitable conclusion that without enforceable mechanisms, such frameworks remain mere positive morality rather than functional safeguards.

(b) Functional Vagueness:

The second model analyzes the semantic instability of ethical principles of AI. As literature indicates, terms like fairness, transparency and others are functionally vague and difficult to operationalize. The *hetu* is the lack of operational thresholds, which is compared to the *drishtanta* of a guideline to "drive properly."

- **Pratijna**: AI ethics frameworks cannot prevent systemic harm.
- **Hetu**: Because they rely on abstract principles that lack operational definitions.
- **Udaharana:** When something is vague, it cannot stop harm.Just as rules that say "drive properly" do not stop accidents.
- **Upanaya:** AI ethics frameworks are functionally vague.
- **Nigamana:** Therefore, AI ethics frameworks cannot stop harm.

This demonstrates that AI ethics frameworks are structurally incapable of governing technical behavior because they rely on abstract principles. Without operational definitions, one cannot regulate complex systems, just as a guideline that merely asks people to drive

properly fails to prevent accidents (Minnesota Department of Transportation, 2010). The idea that generic moral messaging has limited impact on behavioral change has been shown in different domains like health-care (Papakonstantinou et al., 2025) and environmental pollution (Skoric et al., 2022).

(c) Corporate Interests:

The final model deals with the issue of corporate authorship of policies. The *hetu* is that such policies more often than not serve corporate interest rather than public protection.

- **Pratijna**: AI ethics frameworks cannot prevent systemic harm.
- **Hetu**: Because of corporate alignment.
- **Udaharana:** When rules serve corporate interest rather than public interest, they cannot stop harm. Just as safety standards written by manufacturers do not protect consumers.
- **Upanaya:** AI ethics codes are currently self-regulatory.
- **Nigamana:** Therefore, AI ethics frameworks cannot stop harm.

The above *pancha-avayava* demonstrates that AI ethics frameworks are structurally compromised because of corporate capture (*hetu*). The example used is from the history of industrial safety. When safety standards are written solely by manufacturers, they rarely protect consumers because the focus is on cost-effectiveness and not necessarily safety. This is referred to in literature as regulatory capture, a situation where regulatory bodies are controlled by the very industries they are supposed to oversee (Dal Bo, 2006; Kenton, 2025).

Analysis:

The application of the *pancha-avayava* framework shows that, given the three concerns of unenforceability, functional vagueness and corporate capture, contemporary AI ethics frameworks cannot actually stop harm. They are more in the nature of regulatory checklists, good to have on paper but ineffective in terms of governance. This is in fact an example of "ethics washing," where corporations use non-binding principles to preempt laws that would actually threaten their bottom lines (Wagner, 2018).

Specifically in the field of AI ethics, this conclusion is also consistent with emerging critical literature. AI ethics guidelines have been called "meaningless," "useless," and "toothless" since they are incoherent, difficult to apply, and isolated from real-world industry and educational systems. Consequently, they fail to address social, individual, environmental, racial, and structural challenges. Often, they are also actively harmful since they divert resources away from more real meaningful intervention (Mittelstadt et al., 2016; Munn, 2023). Taken together, these critiques provide qualitative support to the findings of the present study.

The study does, however, have some limitations. The analysis focuses on only three structural issues: lack of enforceability, vagueness, and corporate alignment. Additional factors may further strengthen the research hypothesis. Moreover, the establishment of *vyapti* (invariable concomitance) is not demonstrated rigorously since the paper is not empirical in nature. However, the direction of the analysis is correct and future research can build upon this. Empirical testing across real-world AI governance scenarios can add further context and nuance.

Conclusion:

The study uses the *Nyaya pancha-avayava* inferential framework to provide a coherent explanatory account of why existing AI ethics frameworks struggle to prevent harm. Non-enforceable guidelines, functional vagueness, and corporate alignment, stymie the requisite causal mechanisms needed to modify behavior and bring change. In the absence of verifiable empirical examples of non-binding functionally vague corporate-driven principles successfully restraining profit-driven harm, there can be no harm prevention. The study provides a formal logical foundation of ethics-washing grounded in Indian knowledge systems. Until ethics frameworks satisfy the rigors of formal inference and have independent laws and clear consequences, such guidelines remain empty promises.

References:

Corrêa, N. K., Oliveira, N. D., & Massmann, D. (2022). On the efficiency of ethics as a governing tool for artificial intelligence (arXiv:2210.15289). *arXiv*. https://doi.org/10.48550/arXiv.2210.15289

Dal Bó, E. (2006). Regulatory capture: A review. *Oxford Review of Economic Policy, 22*(2), 203–225. https://doi.org/10.1093/oxrep/grj013

Fioravante, R. (2024). Beyond the business case for responsible artificial intelligence: Strategic CSR in light of digital washing and the moral human argument. *Sustainability, 16*(3), 1232. https://doi.org/10.3390/su16031232

Gopinath, K., & Sharma, S. D. (2022). *The computation meme: Computational thinking in the Indic tradition*. IISc Press.

Hagendorff, T. (2020). The ethics of AI ethics: An evaluation of guidelines. *Minds and Machines, 30*(1), 99–120. https://doi.org/10.1007/s11023-020-09517-8

Huang, C., Zhang, Z., Mao, B., & Yao, X. (2023). An overview of artificial intelligence ethics. *IEEE Transactions on Artificial Intelligence, 4*(4), 799–819. https://doi.org/10.1109/TAI.2022.3194503

IBM. (2021, September 17). *What is AI ethics?* https://www.ibm.com/think/topics/ai-ethics

International Organization for Standardization (ISO). (2025). *Building a responsible AI: How to manage the AI ethics debate*. https://www.iso.org/artificial-intelligence/responsible-ai-ethics

Kenton, W. (2025, October 11). Understanding regulatory capture: Definition, impact, and examples. *Investopedia*. https://www.investopedia.com/terms/r/regulatory-capture.asp

Khan, A. A., Akbar, M. A., Fahmideh, M., Liang, P., Waseem, M., Ahmad, A., Niazi, M., & Abrahamsson, P. (2022). AI ethics: An empirical study on the views of practitioners and lawmakers (arXiv:2207.01493). *arXiv*. https://doi.org/10.48550/arXiv.2207.01493

Minnesota Department of Transportation. (2010). *Effectiveness of traffic signs on local roads* (Transportation Research Synthesis Report No. TRS 1002). https://www.lrrb.org/pdf/trs1002.pdf

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2), 2053951716679679. https://doi.org/10.1177/2053951716679679

Mukhopadhyay, S. (2025). AI literacy in the public sphere: A theory-informed exploratory study. *Journal of Artificial Intelligence, Machine Learning and Neural Network, 52*, 24–34. https://doi.org/10.55529/jaimlnn.52.24.34

Munn, L. (2023). The uselessness of AI ethics. *AI and Ethics, 3*(3), 869–877. https://doi.org/10.1007/s43681-022-00209-w

Papakonstantinou, T., Flecke, S. L., Edmunds, C. E. R., Cross, R., Tran, A., & Gold, N. (2025). A systematic review and meta-analysis of the effectiveness of social norms messaging approaches for improving health behaviours in developed countries. *Nature Human Behaviour, 9*(12), 2632–2650. https://doi.org/10.1038/s41562-025-02275-6

Sarukkai, S. (2005). *Indian philosophy and philosophy of science*. Project of History of Indian Science, Philosophy and Culture; Centre for Studies in Civilizations.

Skoric, M. M., Zhang, N., Kasadha, J., Tse, C. H., & Liu, J. (2022). Reducing the use of disposable plastics through public engagement campaigns: An experimental study of the effectiveness of message appeals, modalities, and sources. *International Journal of Environmental Research and Public Health, 19*(14), 8273. https://doi.org/10.3390/ijerph19148273

Wagner, B. (2018). Ethics as an escape from regulation: From ethics-washing to ethics-shopping? In M. Hildebrandt (Ed.), *Being profiling: Cogitas ergo sum* (pp. 108–115). Amsterdam University Press.