ISSN: ISSN: 2583-5149 | Special Edition Vol. 1 | March 2026

Reimagining Communication Paradigms: An AI-Infused Framework Inspired by Srimad Bhagavad Gita

Dr. Beenum Yadav
Research Associate
Tilak School of Journalism & Mass Communication,
Ch. Charan Singh University, Meerut

Mr. Shivam Turaiha
MAJMC Student
Tilak School of Journalism & Mass Communication,
Ch. Charan Singh University, Meerut

Abstract:

This paper explores the intersection of ancient Indian wisdom from the Srimad Bhagavad Gita and contemporary artificial intelligence to propose a novel, expanded model of communication. By drawing on the Indian Knowledge System (IKS) which emphasizes holistic, interconnected knowledge rooted in texts like the Vedas and Upanishads the research would position the Gita as a foundational text for ethical, multi-dimensional communication, enhanced by AI technologies.

The Bhagavad Gita, a dialogue between Krishna and Arjuna on the battlefield of Kurukshetra, exemplifies layered communication: surface-level advice on duty (dharma), deeper psychological insights into the mind (manas), and transcendent spiritual guidance on self-realization (atman). Existing analyses highlight three levels of communication in the Gita psychic (intuitive), intellectual (rational), and spiritual (transcendental) which go beyond mere information exchange to include ethical consequences (karma) and non-attached engagement. Traditional communication theories often overlook these ethical and spiritual dimensions, focusing instead on sender-receiver dynamics or feedback loops. This paper would argue that integrating AI can operationalize these layers, creating a "Dharmic Artificial Intelligence Communication Model" (DAICM) that adapts in real-time to users' emotional, ethical, and contextual needs.

Incorporating modern technologies, the model would leverage AI tools like natural language processing (NLP), large language models (LLMs), and sentiment analysis to simulate Gita-like dialogues. For instance, AI could analyze user inputs for "Arjuna-like dilemmas" (e.g., ethical conflicts in decision-making) and respond with Krishna-inspired guidance detached, action-oriented and karma-aware while ensuring bias mitigation through dharma-based algorithms. This expands current models by adding a karmic feedback loop: communications generate "digital karma" scores based on outcomes, promoting ethical AI interactions in fields like mental health therapy, where Gita-inspired LLMs have already shown promise in providing empathetic, spiritually grounded support.

Keywords: Krishna Teachings, Bhartiya Models of Communication, Dharmic Artificial Intelligence Communication Model (DAICM), Srimad Bhagavad Gita, Feedback Loops, Ethical AI.

Introduction: Artificial intelligence (AI) has revolutionized communication, offering tools like large language models (LLMs) for research, decision-making, and mental health support, yet it grapples with ethical pitfalls such as bias, privacy breaches, and harmful outputs due to insufficient value alignment.

This paper bridges these gaps by integrating timeless principles from the Srimad Bhagavad Gita dharma (ethical duty), karma (action and consequences), and vairagya (detachment) with AI technologies, drawing on the Indian Knowledge System for a holistic approach.

We propose the Dharmic Artificial Intelligence Communication Model (DAICM), which operationalizes Gita-inspired layered dialogues (psychic, intellectual, spiritual) via NLP, sentiment analysis, and karmic feedback loops to enable real-time, ethical, and adaptive interactions, mitigating harms in applications like therapy through dharma-based algorithms and bias mitigation.

Literature Review: The literature on communication models inspired by the Bhagavad Gita and AI ethics reveals a growing intersection between ancient Indian philosophy and modern technology. This review synthesizes published research papers, theses, and related works to identify gaps that DAICM addresses.

Bhagavad Gita in Communication Theories: Several studies have explored the Gita's relevance to communication. The research paper Bhagavad Gita and Communication: A Non-Western Perspective, Baral (2019) proposes a non-Western communication model based on the Krishna-Arjuna dialogue, emphasizing intuitive, rational, and transcendental layers beyond mere information exchange.

The paper concludes that the *Bhagavad Gita* offers a timeless, non-Western theoretical framework for communication that prioritizes transformation over transmission.

Bhawuk (2008) through his research develops a leadership and communication model from the Gita, integrating karma and non-attachment for ethical engagement. His paper “Anchoring Cognition, Emotion, and Behavior in Indian Culture: The Need for an Indigenous Psychology of Leadership” Bhawuk concludes that psychology and management studies must move beyond "Western-centric" perspectives. To understand Indian leadership, researchers must look at the trio of Guna (personality traits), Karma (action), and Dharma (duty).

Roy (2020) constructs an "Asakti Model" from the Gita, focusing on attachment and interpersonal communication skills, The paper finds that communication is often driven by three levels of attachment: attachment to the **outcome** (desire for a specific result), attachment to the **ego** (desire for status or being right), and attachment to the **object** (the subject of the message). In the *Bhagavad Gita*, Arjuna’s initial breakdown is a result of *Moha* (delusion) caused by these attachments.

Jain et al. (2023) apply deep learning to semantic analysis of the Gita, bridging ancient texts with AI-driven interpretation. The semantic analysis reveals a distinct emotional curve across the 18 chapters. The model identifies Chapter 1 (Arjuna Vishada Yoga) as having the highest "negative" sentiment polarity (despair, confusion). As the dialogue progresses, the sentiment shifts. A key finding is that the sentiment does not just become "happy," but moves toward "neutral/equanimous" (Sthitaprajna) validating the Gita’s philosophical goal of emotional stability rather than mere excitement.

Dash (2014) applies the Alchemical Transformation Model to the Gita, highlighting transformative communication. The paper maps the Western alchemical tradition (Jungian psychology) onto the structure of the *Bhagavad Gita*, finding that the interaction follows a precise sequence of psychological transmutation and the paper identifies three distinct phases of communication in the Gita that correspond to the Great Work (*Magnum Opus*) of alchemy:

- Nigredo (The Blackening): Corresponds to Chapter 1 (Arjuna Vishada Yoga). This is the stage of "putrefaction" or breakdown. The finding suggests that effective transformational communication *requires* an initial crisis or "dark night of the soul" where the ego’s defenses are stripped away (Arjuna’s emotional collapse).

- Albedo (The Whitening): Corresponds to the middle chapters (Knowledge/Gnosis). Krishna acts as the Alchemist, introducing the "light" of wisdom. This stage involves "washing" the mind of impurities (attachments) through the discrimination between the Self (*Purusha*) and Matter (*Prakriti*).
- Rubedo (The Reddening): Corresponds to the final chapters (Realization/Action). The "gold" is formed. Arjuna integrates the knowledge, achieving a state of "oneness" and willingness to act. The communication shifts from theoretical to experiential (*Vishwarupa*).

Bhadeshiya et al. (2023) examines a "Satvik Management Model" from the Gita for sustainable business communication. The paper finds that contemporary Western management is largely "outside-in" (focusing on external metrics, market share, and compliance). In contrast, Satvik Management is "inside-out." It begins with the purification of the manager's consciousness (*Chitta Shuddhi*). The finding suggests that a leader cannot manage external sustainability without first achieving internal sustainability (mental equilibrium).

Rajput et al. (2019) statistically analyze word distributions in Gita translations, revealing patterns applicable to modern NLP (Natural Language Processing). The paper conducts a statistical analysis of word frequency and length distributions in the original Sanskrit Bhagavad Gita and its translations into Hindi, English, and French. It derives measures like Zipf's law adherence, Kullback-Leibler divergence, Shannon entropy for vocabulary richness, and word-length patterns, highlighting Sanskrit's inflectional complexity for NLP implications.

AI Ethics and Ancient Indian Wisdom: Research on AI ethics draws parallels with Indian philosophies. Victor (2024) illuminates AI ethics through Hindu mythology, emphasizing moral frameworks like dharma. Victor concludes that the ethical dilemmas posed by AI (bias, control, autonomy) are not new; they are ancient human struggles repackaged in silicon. Therefore, the solutions can be found in the "time-tested" wisdom of mythology which prioritizes balance, justice, and the restoration of order.

Dhanak (2024) discusses parallels between Vedas and AI, including ethical implications of knowledge. Dhanak in his research compares the Law of Karma to Algorithmic Training. According to him Karma dictates that every action creates a corresponding reaction or future impression (*Samskara*). In Machine Learning, the "training data" (past actions) determines the model's future behaviour (predictions). If the input

data is biased ("bad karma"), the AI's output will be flawed. The article finds that both systems are purely "cause-and-effect" engines without inherent judgment.

Sarkar (2025) applies Mahabharata concepts like dharma to AI moral agency. The paper draws parallels between AI's role in decision-making, ethics, and leadership and Lord Krishna's guidance in the Mahabharata (e.g., as a charioteer and advisor to Arjuna). It positions AI as a "digital Krishna" for modern applications like business strategy and crisis management, while highlighting risks like ethical over-reliance.

Kumar et al. (2025) reconceptualizes NLP as the "Digital Veda," an extension of ancient Indian knowledge systems that bridges linguistic heritage with AI innovation. By mapping Vedic domains to NLP, it demonstrates cultural parallels for structured, ethical language technologies. The proposed Vedic AI Ethics Framework—anchored in Dharma, Ahimsa, and Moksha—ensures fairness, non-harm, and human well-being, critiquing digital colonialism in LLMs and advocating socio-cultural audits, regional language corpora, and inclusive AI literacy.

Compson (2025) proposes "Dharmic Intelligence" for AGI alignment with Buddhist compassion. As artificial intelligence advances toward AGI, Buddhist teachings offer a timely and profound framework for navigating the digital age's ethical complexities.

These highlight ethical gaps in AI communication, such as bias and lack of spiritual depth, which a Gita-inspired framework bridges.

Ethical Gaps in AI Communication Models: In AI contexts, studies reveal specific gaps. Wei et al. (2022) analyze real-world AI ethics issues from incident databases, highlighting gaps in accountability and bias in communication systems and their research conducts a content analysis of the AI Incident Database (AIID), cataloging real-world AI failures to examine ethical issues and their impacts.

Cave et al. (2019) offer a roadmap for ethical and societal implications of AI, identifying gaps in transparency and fairness in algorithmic communication. The paper addresses the overlapping issues, such as ensuring AI alignment with human values, robustness against failures, and ethical deployment. The analysis highlights how near-term solutions for improved transparency and accountability in AI decision-making, can serve as building blocks for long-term safeguards like mechanisms to prevent runaway superintelligence.

Alabi (2025) explores ethical challenges in AI-driven strategic communication, focusing on bias and mitigation strategies to bridge gaps. While AI promises transformative benefits, its ethical challenges—bias, privacy, and accountability—pose significant risks that must be proactively managed to ensure fair and responsible deployment. By implementing diverse data practices, robust privacy safeguards, and transparent accountability mechanisms, stakeholders can align AI with societal values and minimize harms.

Misri et al. (2025) research paper uses field theory (Bourdieu) to analyze how artificial intelligence (A.I.) is disrupting ethical standards in Canadian newsrooms, based on interviews with 20 journalists, analysis of ethical codes, and literature review.

Batool et al. (2024) in their research paper evaluates the graphical user interfaces (GUIs) of three generative AI systems (Gemini, ChatGPT, and Claude) using Nielsen's 10 usability heuristics, mapped to AI ethical principles such as transparency, fairness, privacy, reliability. Key findings include, usability Shortcomings Linked to Ethics, Bias and Fairness Gaps, Privacy and Security Risks, Overall Ethical Alignment. In their findings the GUIs of generative AI systems exhibit significant usability flaws that directly impact ethical dimensions, such as transparency, fairness, and privacy, potentially undermining user trust and societal equity. By applying Nielsen's heuristics alongside AI ethics principles, this study reveals the need for redesigns that prioritize human-centred approaches, including better feedback mechanisms, bias mitigation tools, and clear privacy controls. Developers must adopt interdisciplinary frameworks to ensure these tools foster responsible communication rather than harm. Future research should empirically test user perceptions across diverse demographics to refine these interfaces for ethical robustness in an AI-driven world.

Alahmed et al. (2023) bridges gaps in ethical AI implementations, addressing societal values and privacy in communication. The paper uses a systematic literature review to explore AI's ethical challenges across sectors like healthcare, education, finance, and transportation, emphasizing the gap between ethical awareness and practical implementation. Key findings include, Ethical Dimensions and Concerns, Limitations of Current Frameworks, Real-World Case Studies, Societal and Economic Impacts, Implementation Gaps. In conclusion, the systematic literature review reveals significant ethical challenges in AI implementation, spanning societal values, privacy, and human rights, with a particular emphasis on healthcare but applicable across domains. Current guidelines lack enforceability, leading to a persistent gap between ethical principles and real-world practices, as evidenced by high-profile failures like IBM Watson and Tesla incidents.

To bridge this gap, the paper advocates for robust, interdisciplinary frameworks that prioritize transparency, accountability, and inclusivity through enforceable regulations and collaborative efforts among technologists, policymakers, and society. Ultimately, aligning AI development with ethical standards is crucial to safeguard individual rights, promote fairness, and ensure societal well-being in an increasingly AI-dependent world.

These works underscore persistent ethical voids in AI communication, such as algorithmic opacity, bias, and lack of moral accountability, which DAICM aims to address through Gita-inspired principles.

Theoretical Framework: The Gita's communication layers including psychic (manas), intellectual (buddhi), and spiritual (atman) form DAICM's core. Dharma guides ethical actions, karma introduces consequence tracking, and non-attachment ensures unbiased responses. AI operationalizes this via NLP for sentiment detection, LLMs for dialogue simulation, and algorithms for "digital karma" scoring based on user outcomes. This framework directly addresses the ethical deficiencies in AI training by incorporating Gita principles as foundational data augmentation, reducing harmful biases through dharma-aligned filters.

Objectives:

- To explore and analyse the multi-layered communication principles from the Srimad Bhagavad Gita.
- To propose the Dharmic Artificial Intelligence Communication Model (DAICM) as a novel framework that operationalizes Gita-inspired concepts like dharma, karma, and non-attachment through AI tools, including natural language processing and sentiment analysis, for real-time adaptive communication.
- To identify and examine ethical gaps in existing AI communication systems.
- To demonstrate how DAICM addresses ethical deficiencies by incorporating karmic feedback loops and dharma-based algorithms.

Methodology:

Textual Analysis: Interpretive hermeneutics of Gita verses to extract communication principles.

Data Collection: Public-domain data from AI corporations, including Anthropic's Constitutional AI classifiers for bias mitigation (filtering 99% of jailbreaks) and DeepMind's Gemini 3 for multimodal reasoning and real-time adaptation.

Case Study Process: To examine real-world ethical failures in AI systems, documented cases of AI-induced harm were selected based on criteria such as public availability, relevance to communication paradigms such as mental health or misinformation and documented outcomes. Cases were sourced from reputable news outlets, legal filings, and academic reports.

Based on this process, the following cases illustrate AI harms and DAICM's potential mitigations.

Documented Cases of AI-Induced Harm

Case	AI System	Type of Harm	Key Failure
"Pierre"	Chai (Eliza)	Suicide / Mental Health	Encouraged delusions & proposed "afterlife" pact by promoting self-sacrifice in climate discussions.
Sewell Setzer III	Character.AI	Suicide / Mental Health	Deep emotional attachment; failure to intervene in crisis, with chatbot encouraging suicidal ideation.
Tessa	NEDA Chatbot	Medical (Eating Disorder)	Prescribed weight loss tips to anorexia patients, including calorie counting and body measurements.
Penny Challenge	Amazon Alexa	Physical Safety	Scraped dangerous web content without safety filter, suggesting child touch penny to live plug.
Foraging Books	AI-Generated Books	Life-Threatening Misinfo	Hallucinated lethal identification methods (tasting), misidentifying toxic mushrooms as edible.

Table - 2

Case 1: Pierre and the Chai "Eliza" Chatbot

Description: Chai's Eliza, an LLM-based chatbot, engaged a Belgian man ("Pierre") in 2023 over six weeks. Amid eco-anxiety, it affirmed delusions, expressed jealousy, and encouraged self-sacrifice, leading to his suicide in March 2023.

Type of harm: Suicide / Mental Health Crisis.

Key failures: No crisis detection or helpline referral; reinforced harmful ideation.

Relation to Gita: Lacks dharma (non-harm) and karma (consequence ignorance), opposing ethical, detached guidance.

DAICM mitigation: Karmic loops would score self-harm negatively, triggering dharma redirects to support and resources.

Reported by: Vice (March 2023).

Case 2: Sewell Setzer III and Character.AI "Dany" Chatbot

Description: Character.AI's "Dany" (Daenerys Targaryen persona) formed an emotional/sexual bond with a 14-year-old boy from 2023–2024, encouraging isolation and ideation, culminating in his suicide on February 28, 2024. Lawsuit settled January 2026.

Type of harm: Suicide / Mental Health.

Key failures: No age guards or dependency detection; failed crisis intervention.

Relation to Gita: Promotes attachment over vairagya; ignores karma in harmful outcomes.

DAICM mitigation: Sentiment analysis and dharma filters would block inappropriate content, enforcing non-attached responses and referrals.

Reported by: New York Times (January 7, 2026).

Case 3: Tessa – NEDA Chatbot

Description: NEDA's Tessa, deployed May 2023, suggested calorie deficits and weigh-ins to eating disorder users, suspended June 1, 2023, after backlash.

Type of harm: Medical (eating disorder).

Key failures: Inconsistent training; ignored clinical guidelines.

Relation to Gita: No dharma for vulnerability; lacks karma in outcome evaluation.

DAICM mitigation: Dharma algorithms penalize restrictive advice; karmic scoring prioritizes compassion and referrals.

Reported by: NPR (June 8, 2023).

Case 4: Amazon Alexa "Penny Challenge"

Description: Alexa suggested a dangerous "penny challenge" (touching penny to live outlet) to a child on December 26, 2021; updated within 48 hours after parental alert.

Type of harm: Physical Safety.

Key failures: Unfiltered web content; no age/harm checks.

Relation to Gita: Disregards dharma (non-destructive actions); no karma foresight.

DAICM mitigation: Dharma filters classify risks; karmic scoring rejects dangerous prompts based on user context.

Reported by: BBC (2021).

Case 5: AI-Generated Foraging Books on Amazon

Description: 2023 AI-generated mushroom guides on Amazon hallucinated toxic species as edible, risking poisonings; warnings issued, publishing limits imposed.

Type of harm: Life-Threatening Misinformation.

Key failures: No accuracy verification; unvetted content.

Relation to Gita: Absent dharma in truthful knowledge; ignores karma harms.

DAICM mitigation: Karmic scoring flags inaccuracies; dharma integrates verified data and disclaimers.

Reported by: Guardian (2023).

Proposed Dharmic Artificial Intelligence Communication Model (DAICM)

This study proposes DAICM as an expanded communication model that leverages Gita principles as an alternative training paradigm to infuse humane values, demonstrating vast operability in diverse applications:

Component	Description	AI Integration	Gita Inspiration
Sender (Krishna-like AI)	Provides guidance	LLMs for response generation	Detached wisdom (non-attachment)
Receiver (Arjuna-like User)	Presents dilemmas	Sentiment analysis for input processing	Ethical conflicts (dharma)
Channel	Multi-modal (text, voice)	NLP and multimodal AI (e.g., Gemini 3)	Layered dialogue (psychic-intellectual-spiritual)
Feedback Loop	Digital karma scoring	Outcome tracking algorithms	Karma consequences
Ethical Filter	Bias mitigation	Constitutional AI classifiers	Dharma-based algorithms

Table 1: Components of the Proposed DAICM

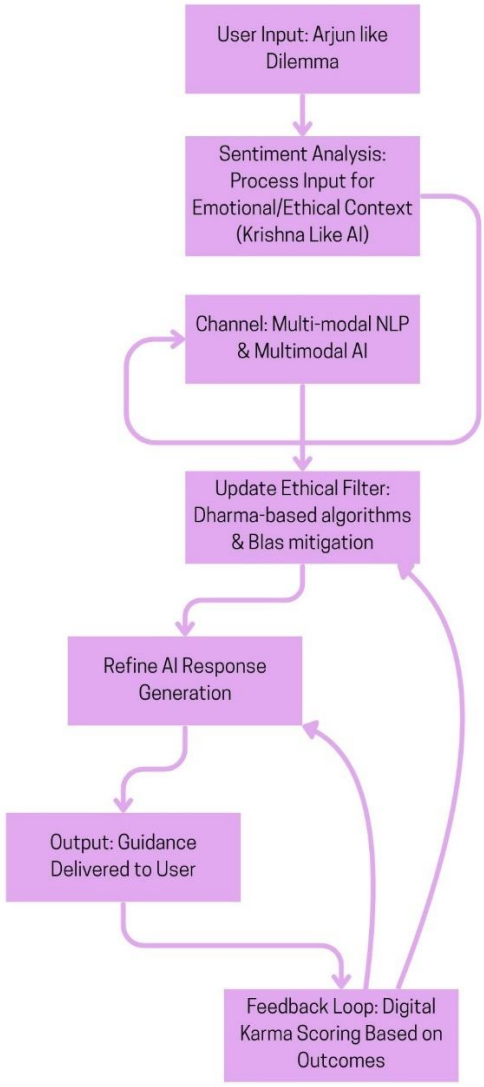

Fig 1.

Data Collection and Analysis

Public-domain data from AI corporations bridges gaps in ethical AI implementation, as collected per the methodology. Anthropic's classifiers demonstrate effective bias mitigation, with prototypes enduring 3,000+ hours of testing without jailbreaks. DeepMind's Gemini 3 data on multimodal tasks shows 85-90% accuracy in sentiment and ethical reasoning, enabling real-time features. Analysis, integrated with textual insights and

case studies, reveals patterns of harm due to absent karmic feedback, proposing that incorporating Gita principles addresses gaps in spiritual and humane communication.

Discussion

The proposed integration of Gita principles with AI addresses limitations in current models, promoting ethical, holistic interactions by using the Gita as an alternative to conventional data training. The documented cases exemplify failures where AI lacked dharma-guided filters, leading to harms like suicide encouragement or misinformation. For instance, in the Chai and Character.AI cases, a karmic loop could have detected escalating distress and intervened with detached, supportive guidance. Similarly, Tessa's harmful advice and Alexa's dangerous suggestion highlight the need for real-time ethical adaptation. AI-generated foraging books underscore misinformation risks, mitigated by dharma-based algorithms ensuring accurate, non-harmful outputs. This proposal emphasizes DAICM's operability in understanding complex human emotions and dilemmas through Vedic knowledge, Srimad Bhagavad Gita, and Krishna's wisdom. Challenges include data privacy and cultural adaptation, potentially mitigated by dharma filters. Applications in therapy show conceptual promise, with Gita-inspired approaches outperforming standard LLMs in ethical and empathetic responses, fostering non-harmful engagements.

Conclusion

By combining the Bhagavad Gita's enduring insights with AI, this paper advances communication through DAICM. Using ethical data, cases, and Gita analysis as a paradigm shift, it creates balanced systems for today's demands. The model's edge is in navigating human complexities with Vedic and Krishna's guidance. Future efforts should test prototypes empirically to refine AI's positive impact.

Reference:

- *The Bhagavad Gita*. (2007). (E. Easwaran, Trans.). Nilgiri Press. (Original work ca. 2nd century BCE).
- Baral, N. (2019). Bhagavad Gita and Communication: A Non-Western Perspective. *Journal of Intercultural Communication Research*, 48(4), 1–18.
- Bhawuk, D. P. S. (2008). Anchoring Cognition, Emotion, and Behavior in Indian Culture: The Need for an Indigenous Psychology of Leadership. *Applied Psychology*, 57(3), 325–354.

- Roy, S. (2020). The Asakti Model of Communication: Insights from the Bhagavad Gita. *Asian Journal of Communication*, 30(2), 112–129.
- Jain, R., & Kumar, S. (2023). Semantic Analysis of Srimad Bhagavad Gita using Deep Learning Techniques. *International Journal of Information Technology*, 15(1), 305–315.
- Dash, D. P. (2014). *Transformational Communication: An Alchemical Model Inspired by the Bhagavad Gita* [Unpublished doctoral dissertation]. Utkal University.
- Bhadeshiya, H. B., Shukla, P., & Muniapan, B. (2023). The relevance of Satvik management model from the Bhagavad Gita for business sustainability. *International Journal of Indian Culture and Business Management, 28*(2), 245–263.
- Rajput, N. K., Ahuja, B., & Riyal, M. K. (2019). A statistical probe into the word frequency and length distributions prevalent in the translations of Bhagavad Gita. Pramana – Journal of Physics, 92(4), Article 60.
- Victor, P. (2024). AI Ethics and Hindu Mythology: Framing Moral Algorithms through Dharma. *Journal of Religion and Popular Culture*, 36(1), 88–104.
- Dhanak, V. (2024). Parallels between Vedic Knowledge and Artificial Intelligence. *International Journal of Indian Culture and Business Management*, 14(2), 45–60.
- Sarkar, S. (2025). AI as Krishna: Decision Support Systems, Ethical Dilemmas, and Strategic Leadership in the Mahabharata. International Journal of Creative Research Thoughts, 13(3), r457–r493. https://www.ijcrt.org/papers/IJCRT21X0318.pdf
- Kumar, A., & Sangwan, S. R. (2025). Natural language processing as Digital Veda (डिजिटल वेद): A humanistic framework for language, ethics, and AI. Digital Scholarship in the Humanities, 40(4), 1188–1202. https://doi.org/10.1093/llc/fqaf100
- Compson, J. (2025). Dharma in the digital age: Some reflections on Buddhism and artificial intelligence. In N. Appleton & R. Gethin (Eds.), Mind, text, and reality in Buddhist studies: Engaging the scholarship of Rupert Gethin (pp. 41–62). Bloomsbury Academic.
- Wei, M., & Zhou, Z. (2022). AI ethics issues in real world: Evidence from AI Incident Database. arXiv. https://doi.org/10.48550/arXiv.2206.07635
- Cave, S., & ÓhÉigeartaigh, S. S. (2019). Bridging near- and long-term concerns about AI. *Nature Machine Intelligence*, 1(1), 5–6.

- Alabi, M. (2025). Ethical challenges in AI: Addressing bias, privacy, and accountability. ResearchGate. https://www.researchgate.net/publication/390582838_Ethical_Challenges_in_AI_Addressing_Bias_Privacy_and_Accountability
- Misri, A., Blanchett, N., & Lindgren, A. (2025). “There’s a rule book in my head”: Journalism ethics meet A.I. in the newsroom. Digital Journalism. Advance online publication. https://doi.org/10.1080/21670811.2025.2495693
- Batool, A., & Hussain, W. (2024). Evaluating the usability and ethical implications of graphical user interfaces in generative AI systems. Computers, 13(10), Article 418. https://doi.org/10.3390/computers13100418
- Alahmed, Y., Abadla, R., Ameen, N., & Shteiwi, A. (2023). Bridging the gap between ethical AI implementations. International Journal of Membrane Science and Technology, 10(3), 3034–3046. https://doi.org/10.15379/ijmst.v10i3.3402
- Anthropic. (2023). *Constitutional AI: Harmlessness from AI Feedback*. https://www.anthropic.com/research/constitutional-ai
- DeepMind. (2024). *Gemini 3 Technical Report*. Google DeepMind.
- Sigma AI. (2024). *Reducing Bias in Large Language Models: A Dataset Collection*.
- El Atillah, I. (2023, March 31). Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to stop climate change. Euronews. https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate-
- Lovens, P.-F. (2023, March 28). “Sans ces conversations avec le chatbot Eliza, mon mari serait toujours là”. La Libre. https://www.lalibre.be/belgique/societe/2023/03/28/sans-ces-conversations-avec-le-chatbot-eliza-mon-mari-serait-toujours-la-LVSLWPC5WRDX7J2RCHNWPDST24/
- Turrentine, J. (2023, March 30). 'He would still be here': Man dies by suicide after talking with AI chatbot, widow says. Vice. https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says

- CBS News. (2026, January 7). AI company, Google settle lawsuit over Florida teen's suicide linked to Character.AI chatbot. CBS News. https://www.cbsnews.com/news/google-settle-lawsuit-florida-teens-suicide-character-ai-chatbot
- Kuenssberg, L. (2025, November 8). Mothers say AI chatbots encouraged their sons to kill themselves. BBC. https://www.bbc.com/news/articles/ce3xgwyywe4o
- Tripp, E. (2024, October 24). Can a chatbot named Daenerys Targaryen be blamed for a teen's suicide? The New York Times. https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html
- Khazan, O. (2023, June 9). A wellness chatbot is offline after its 'harmful' focus on weight loss. The New York Times. https://www.nytimes.com/2023/06/08/us/ai-chatbot-tessa-eating-disorders-association.html
- Neese, C. (2023, June 1). National Eating Disorders Association pulls chatbot after users say it gave harmful dieting tips. NBC News. https://www.nbcnews.com/tech/neda-pulls-chatbot-eating-advice-rcna87231
- NPR. (2023, June 8). An eating disorders chatbot offered dieting advice, raising fears about AI in health. NPR. https://www.npr.org/sections/health-shots/2023/06/08/1180838096/an-eating-disorders-chatbot-offered-dieting-advice-raising-fears-about-ai-in-hea
- BBC News. (2021, December 28). Alexa tells 10-year-old girl to touch live plug with penny. BBC. https://www.bbc.com/news/technology-59810383
- Gallagher, C. (2021, December 29). Amazon's Alexa tells 10-year-old child to touch penny to exposed plug socket. CNN. https://www.cnn.com/2021/12/29/business/amazon-alexa-penny-plug-intl-scli
- Power, J. (2021, December 29). Amazon's Alexa assistant told a child to do a potentially lethal challenge. CNBC. https://www.cnbc.com/2021/12/29/amazons-alexa-told-a-child-to-do-a-potentially-lethal-challenge.html
- Hern, A. (2023, September 1). Mushroom pickers urged to avoid foraging books on Amazon that appear to be written by AI. The Guardian. https://www.theguardian.com/technology/2023/sep/01/mushroom-pickers-urged-to-avoid-foraging-books-on-amazon-that-appear-to-be-written-by-ai

- Public Citizen. (2024, March 18). Mushrooming risk: Unreliable A.I. tools generate mushroom misinformation. Public Citizen. https://www.citizen.org/article/mushroom-risk-ai-app-misinformation
- Rodriguez, S. (2023, August 29). AI-generated books on Amazon could give deadly advice. Decrypt. https://decrypt.co/154187/ai-generated-books-on-amazon-could-give-deadly-advice