Risks in Multi-domain AI Assistants: A Literature-Based Analysis with Focus on the Indian Context

Yogeshwari Yawalkar, Mahek Jameer Pathan, Misba Shakil Khan, Pravin Sakharam Barela

D.Y. Patil Institute of Master of Computer Applications and Management, Akurdi, Pune.
Students, Department of Master of Computer Applications

Abstract

Multi-domain artificial intelligence (AI) assistants—systems capable of handling diverse tasks spanning conversation, code generation, information retrieval, planning, and tool use—have rapidly become integral to personal, enterprise, and public-sector workflows. Their versatility expands the attack surface and amplifies risks that single-purpose models do not fully exhibit. This literature-based research paper synthesizes findings from more than twenty peer-reviewed studies, industry frameworks, and policy documents published primarily between 2021 and 2026. It examines technical risks (prompt injection, excessive agency, hallucination, data leakage), socio-technical risks (bias, misinformation, representational harms), and systemic risks (supply-chain vulnerabilities, multi-agent failure modes). Special attention is given to the Indian context, including the implications of the Digital Personal Data Protection (DPDP) Act, 2023, linguistic and cultural diversity challenges, digital public infrastructure exposure, and emerging governance gaps. A comparative review of 22 key works is presented. The paper concludes with recommendations for layered technical controls, organizational practices, and policy interventions suited to India's regulatory and socio-economic landscape.

Keywords: multi-domain AI assistants, large language models, agentic AI, prompt injection, hallucination, bias, privacy, DPDP Act, India, AI governance, systemic risk

Introduction

The transition of large language models (LLMs) from specialized research tools to general-purpose, multi-domain assistants marks one of the most consequential shifts in contemporary computing. Systems such as ChatGPT, Claude, Gemini, Grok, and a growing ecosystem of open-weight and enterprise agents now routinely perform tasks that previously required separate applications: drafting documents, writing and debugging code, summarizing research, retrieving and synthesizing web information, planning multi-step workflows, and interacting with external tools and application programming interfaces (APIs). This multi-domain capability confers

substantial productivity benefits but simultaneously concentrates risk. A single compromised or misbehaving assistant can affect confidentiality, integrity, and availability across multiple functional domains and organizational boundaries.

Scholarly and industry attention has correspondingly intensified. Early work focused on intrinsic model limitations—hallucination, bias, and toxicity [6], [7]. Subsequent research documented the practical exploitability of instruction-following models through direct and indirect prompt injection [1], [2]. More recent systematizations have turned to agentic architectures, where models plan, use tools, maintain memory, and coordinate with other agents, revealing qualitatively new failure modes such as excessive agency, inter-agent infection, and cross-principal confidentiality breaches [13], [14], [15]. Industry frameworks, most notably the OWASP Top 10 for Large Language Model Applications [9], have crystallized practitioner consensus around the highest-priority risks.

India occupies a distinctive position in this landscape. The country accounts for a disproportionately large share of global AI chatbot downloads and is rapidly integrating AI into digital public infrastructure, financial services, healthcare, and governance. At the same time, its primary data-protection statute—the Digital Personal Data Protection Act, 2023—together with evolving AI governance guidelines, must address risks that were only partially anticipated when the legislation was drafted. Linguistic diversity, varying digital literacy, and the scale of public-sector deployments further shape the risk profile [21], [22].

Problem Statement

Despite the rapid proliferation of multi-domain AI assistants, existing research remains fragmented across technical security studies, bias and fairness literature, agentic system analyses, and policy documents. There is limited integrated understanding of how these risks interact when a single assistant operates across multiple domains, and even less attention has been paid to how these risks manifest in the Indian regulatory, linguistic, and infrastructural context. Without a coherent synthesis, researchers, practitioners, and policymakers lack a clear map of priority risks and context-sensitive mitigation pathways.

Objectives

This literature-based study pursues the following objectives:

1. To synthesize and organize the major categories of risk associated with multi-domain AI assistants as documented in peer-reviewed literature, industry frameworks, and policy sources between 2021 and 2026.
2. To present a comparative overview of key research works, highlighting their focus, methodology, and relevance to multi-domain assistants.
3. To examine how these risks specifically apply within the Indian context, with particular attention to the DPDP Act, linguistic diversity, digital public infrastructure, and emerging governance gaps.
4. To identify practical mitigation principles and open challenges that can guide future research, system design, and policy development.

Literature Review

The literature on risks in multi-domain AI assistants has evolved rapidly. Early foundational work examined intrinsic limitations of large language models. Bender et al. [6] cautioned against the uncritical scaling of language models, highlighting bias amplification, environmental costs, and opacity. Weidinger et al. [7] offered one of the first comprehensive taxonomies of risks from language models, covering discrimination, information hazards, misinformation, malicious uses, human–computer interaction harms, and broader socioeconomic effects. Bommasani et al. [8] expanded the discussion to foundation models more generally, emphasizing homogenization, dual-use potential, and concentration of power.

Privacy risks received early empirical attention. Carlini et al. [5] demonstrated that large language models can memorize and regurgitate training data, including personally identifiable information. Subsequent work on extraction attacks reinforced that sensitive information disclosure remains a persistent concern [9].

From 2023 onward, research shifted toward the interaction layer. Greshake et al. [1] showed that real-world LLM-integrated applications are vulnerable to indirect prompt injection, in which adversarial instructions hidden in external content (web pages, documents, emails) can hijack model behavior. Liu et al. [2] and Wei et al. [3] systematically studied jailbreaking, revealing that safety training can fail when models face competing objectives or distributional shift. Zou et

al. [10] demonstrated that adversarial suffixes can transfer across models, underscoring the difficulty of achieving robust alignment.

Hallucination has been extensively surveyed. Huang et al. [11] and later reviews document high rates of factual and faithfulness failures, noting that retrieval-augmented generation mitigates but does not eliminate the problem. Bias and fairness research has similarly matured; Gallegos et al. [12] provide a detailed survey of social bias evaluation and mitigation techniques for LLMs.

The rise of agentic systems has generated a new wave of security literature. Yao et al. [13] and Deng et al. [14] survey threats specific to AI agents, including tool misuse, multi-step attacks, and multi-agent cascade failures. Ko et al. [15] identify seven novel security challenges that arise when agents collaborate across organizational trust boundaries. Li et al. [16] extend the analysis to multimodal large language models, highlighting cross-modal attack surfaces. At the systemic level, Uuk et al. [17] propose a taxonomy of large-scale societal risks from general-purpose AI.

Industry frameworks have paralleled academic work. The OWASP Top 10 for LLM Applications [9] currently ranks prompt injection, sensitive information disclosure, and excessive agency among the highest-priority risks. Foundational safety papers by Amodei et al. [18] and Hendrycks et al. [19] remain relevant, particularly their discussions of reward hacking, scalable oversight, and distributional shift. Shelby et al. [20] contribute a sociotechnical harm taxonomy that helps connect technical failures to human impacts.

Indian-specific sources remain fewer in the peer-reviewed literature but are growing in policy and industry reporting. The Ministry of Electronics and Information Technology's AI Governance Guidelines and related white papers [21] flag fragmented oversight, weak training-data governance, and the difficulty of correcting unsafe behavior after large-scale deployment. An industry survey by Delinea [22] reports that a high proportion of Indian enterprises have experienced AI agents accessing sensitive data beyond intended scope, despite formal policies.

Comparative Review of Key Studies

To provide a clearer picture of the existing body of work, the following comparative review summarizes 22 key studies and documents. Each entry highlights the primary focus, the main

risks identified, the methodological approach, and the relevance of the work to multi-domain AI assistants.

1. Greshake et al. (2023) examined indirect prompt injection attacks on real-world LLM-integrated applications. Using empirical attack demonstrations, they showed how adversarial instructions hidden in external content can compromise system behavior. This work is highly relevant because multi-domain assistants routinely process untrusted retrieved content and tool outputs.
2. Liu et al. (2023) conducted an empirical study of jailbreaking ChatGPT through prompt engineering. Their findings revealed that alignment mechanisms can be bypassed to produce harmful outputs, underscoring persistent safety challenges for general-purpose assistants.
3. Wei et al. (2023) analyzed why LLM safety training fails, identifying competing objectives and mismatched generalization as core problems. Their theoretical and empirical analysis remains foundational for understanding safety failures in multi-domain systems.
4. Perez et al. (2022) pioneered red teaming of language models using both human and automated methods. They documented risks related to toxic, biased, and privacy-leaking outputs, providing an early multi-risk taxonomy still useful today.
5. Carlini et al. (2021) demonstrated that large language models can memorize and leak training data through extraction attacks. Their work established sensitive information disclosure as a core privacy risk for any assistant trained on large corpora.
6. Bender et al. (2021), in their widely cited “stochastic parrots” paper, critically examined the societal risks of large language models, including bias amplification, environmental cost, and opacity. The analysis remains central to discussions of multi-domain assistants.
7. Weidinger et al. (2022) proposed a comprehensive taxonomy of risks posed by language models, covering discrimination, misinformation, and human–computer interaction harms. Their literature synthesis continues to serve as a broad organizational framework.

8. Bommasani et al. (2021) surveyed the opportunities and risks of foundation models, highlighting homogenization, dual-use potential, and concentration of power. Their position paper remains essential for understanding systemic multi-domain risks.
9. The OWASP GenAI Security Project (2025/2026) published the Top 10 for Large Language Model Applications. Through community consensus, it ranks prompt injection, sensitive information disclosure, excessive agency, and related risks, serving as the leading industry-standard framework.
10. Zou et al. (2023) developed universal and transferable adversarial attacks on aligned language models using gradient-based optimization. Their work demonstrated that jailbreaks can transfer across models, raising concerns about robustness in multi-domain deployments.
11. Huang et al. (2023) conducted a systematic survey of hallucination in large language models, examining factuality and faithfulness failures. Their review is directly relevant to the reliability of multi-domain assistants across knowledge-intensive tasks.
12. Gallegos et al. (2024) provided a comprehensive survey of bias and fairness in large language models, covering social bias, stereotyping, and unequal performance across groups. Their work informs fairness considerations for diverse user populations.
13. Yao et al. (2024) surveyed security and privacy issues in large language models and early agentic systems, focusing on tool misuse and multi-step attacks. Their taxonomy is particularly useful for understanding risks in agentic multi-domain assistants.
14. Deng et al. (2024) offered a systematic survey of security challenges facing AI agents, organizing threats around perception, planning, action, and multi-agent interactions. The paper highlights important knowledge gaps in agent security research.
15. Ko et al. (2026) mapped seven novel security challenges that arise in cross-domain multi-agent LLM systems. Their position paper is especially relevant for assistants that collaborate across organizational or trust boundaries.
16. Li et al. (2026) surveyed the evolving safety landscape of multi-modal large language models, identifying cross-modal attacks and modality misalignment as emerging threats. This work extends risk analysis to multimodal multi-domain assistants.

17. Uuk et al. (2024) developed a taxonomy of systemic risks from general-purpose AI based on a systematic review of 86 papers. They identified 13 categories of large-scale societal risks, offering a macro-level perspective relevant to widespread multi-domain deployment.
18. Amodei et al. (2016) outlined concrete problems in AI safety, including reward hacking, scalable oversight, and distributional shift. Although written before the current generation of assistants, their analysis remains foundational.
19. Hendrycks et al. (2021) set out unsolved problems in machine learning safety, covering robustness, monitoring, alignment, and systemic risks. Their agenda-setting paper continues to frame long-term concerns for multi-domain systems.
20. Shelby et al. (2023) proposed a taxonomy of sociotechnical harms of algorithmic systems, including representational, allocative, and quality-of-service harms. Their framework helps connect technical failures of assistants to concrete human impacts.
21. The Ministry of Electronics and Information Technology (MeitY) and the Office of the Principal Scientific Adviser (2025–2026) issued AI Governance Guidelines and a related white paper. These policy documents highlight bias, data misuse, deepfakes, opacity, and the challenges of post-facto enforcement in the Indian context.
22. Delinea (2026) conducted an industry survey focused on Indian enterprises and found that a large majority reported AI agents accessing sensitive data beyond intended scope, even though formal AI policies were in place. The survey provides empirical evidence of the enforcement gap in the Indian enterprise setting.

Taken together, these works show a clear progression: early literature concentrated on intrinsic model properties such as bias and hallucination, while later research increasingly addresses interaction-layer and system-level threats arising from tool use, memory, and multi-agent coordination. Explicit treatment of the Indian regulatory and socio-technical context remains limited within the academic corpus, reinforcing the need for the contextual analysis presented in this paper.

Methodology

This study adopts a literature-based, qualitative synthesis approach. No original experiments or primary data collection were conducted. Instead, the analysis draws on peer-reviewed journal and conference papers, high-impact preprints, industry frameworks, and official policy documents published mainly between 2021 and 2026.

Sources were identified through targeted searches of academic databases and preprint repositories (including arXiv), citation tracking from key papers, and review of official Indian government and industry publications. Inclusion criteria prioritized works that (a) address risks associated with large language models, foundation models, or agentic AI systems; (b) offer taxonomies, empirical findings, or systematic surveys; and (c) contribute either technical, socio-technical, or policy perspectives. Indian policy documents and recent industry surveys were deliberately included to ground the analysis in the local regulatory and deployment context.

The selected literature was organized thematically into risk categories (technical, socio-technical, and systemic) and then examined for relevance to multi-domain assistants. Findings were synthesized narratively, with special attention to interactions among risks and to implications for the Indian setting under the DPDP Act and related governance instruments.

Results and Discussion

Major Risk Categories

The synthesis of the literature yields seven interrelated risk categories that are particularly salient for multi-domain AI assistants.

Prompt Injection and Instruction Hijacking

Prompt injection remains the most consistently ranked risk across industry frameworks [9]. In multi-domain assistants the threat is amplified because the model processes both user instructions and untrusted external content—retrieved documents, web pages, emails, or tool outputs—within a shared context window. Indirect prompt injection can lead to unauthorized tool invocations, data exfiltration, or persistent behavioral changes if the system maintains long-term memory [1]. Defenses that rely solely on the model's ability to distinguish instructions from data have repeatedly proven insufficient against adaptive adversaries.

Excessive Agency and Autonomous Action

As assistants gain the ability to call tools, execute code, send messages, or modify external state, the consequences of a successful injection or alignment failure escalate from textual harm to real-world impact. Empirical studies of agentic systems show that refusal rates for harmful requests can drop substantially once tools are available [13], [14]. In multi-agent settings, a single compromised agent can propagate malicious instructions or poisoned information to peers, creating cascade failures that local monitoring may miss [15].

Hallucination and Misinformation

Hallucination undermines reliability in any domain that requires accuracy. Multi-domain assistants exacerbate the problem because users may not know which of the assistant's many capabilities are more or less prone to fabrication. Systematic reviews document high rates of hallucination in knowledge-intensive tasks and note that retrieval-augmented generation mitigates but does not eliminate the risk [11]. In high-stakes Indian contexts such as legal research, medical advice, or welfare eligibility, hallucinated outputs can produce tangible harm.

Privacy, Data Leakage, and Sensitive Information Disclosure

Multi-domain assistants often operate with access to personal data, enterprise documents, conversation history, and retrieved external content. Risks include memorization and extraction of training data [5], inadvertent disclosure of system prompts or other users' information, cross-tenant leakage, and the generation of hallucinated but personally identifiable information about real individuals. Under data-protection regimes that emphasize purpose limitation and consent, the dynamic nature of agentic workflows complicates compliance demonstration.

Bias, Fairness, and Representational Harms

Foundation models trained on internet-scale data inherit and can amplify social stereotypes and unequal performance across demographic groups and languages [6], [12]. In a multi-domain assistant, bias may manifest differently depending on the task—gender stereotyping in career advice, differential accuracy in medical information, or poorer performance on non-English or low-resource Indian languages. Representational harms accumulate over repeated interactions and can reinforce social inequities even when no single output appears egregiously discriminatory.

Supply-Chain, Poisoning, and Ecosystem Risks

Multi-domain assistants depend on complex supply chains: base models, fine-tuning datasets, embedding models, vector databases, tool plugins, and orchestration frameworks. Compromises at any layer can introduce backdoors or degrade safety properties [9]. The increasing use of retrieval from the open web further expands the attack surface for knowledge-base poisoning.

Multi-Agent and Cross-Domain Systemic Risks

When multiple agents collaborate across organizational or trust boundaries, novel failure modes emerge: miscoordination, conflict, collusion, containment failure, and transitive authorization failures [15], [17]. These risks are particularly salient for multi-domain assistants that may themselves spawn or interact with specialized sub-agents.

The Indian Context

India's primary statutory instrument governing personal data is the Digital Personal Data Protection Act, 2023 (DPDP Act), supplemented by the DPDP Rules. The Act is consent-centric and imposes obligations of purpose limitation, data minimization, reasonable security safeguards, and breach notification, with penalties that can reach ₹250 crore. Applying these obligations to multi-domain AI assistants raises operational challenges: agentic systems may decide at runtime which data to retrieve or infer, making static purpose statements difficult to maintain; hallucinated personal data can trigger liabilities; and cross-border model providers introduce questions of transfer and localization [21].

Beyond data protection, India's AI Governance Guidelines emphasize innovation, accountability, safety, and inclusion while relying largely on existing laws rather than a comprehensive horizontal AI statute. Policy analyses have flagged fragmented and reactive oversight, weak data governance at the training stage, and unresolved liability for actions taken by autonomous agents under the Information Technology Act, 2000 [21].

Several socio-technical factors further shape the risk profile. Linguistic and cultural diversity means that models trained predominantly on English may exhibit higher hallucination rates or culturally inappropriate outputs when serving users in Indian languages. Digital literacy varies widely, increasing the potential harm of fluent but incorrect advice. Integration—direct or indirect—with digital public infrastructure (Aadhaar, UPI, direct benefit transfers) concentrates systemic stakes. Industry surveys indicate that Indian organizations report high rates of AI agents

accessing sensitive data beyond intended scope despite formal policies [22]. Mental-health and emotional-support uses of chatbots raise additional concerns that existing content-moderation and medical-device frameworks do not fully address.

Mitigation Principles and Open Challenges

The literature converges on several practical principles. Assume that prompt injection and related attacks will occasionally succeed; therefore constrain the blast radius through least-privilege permissions, human-in-the-loop approval for high-impact actions, and robust output handling. Treat external content and tool outputs as untrusted. Invest in continuous red-teaming that covers multi-step, multi-modal, and multi-agent scenarios. For privacy, minimize sensitive data in context windows, apply retrieval access controls, and maintain audit logs. Address bias through data curation, evaluation on diverse Indian language and demographic benchmarks, and ongoing monitoring of real-world outputs.

Open challenges remain substantial. Model-internal safety training often degrades under tool use and long-horizon interaction. Multi-agent trust and containment mechanisms lag behind the pace of deployment. Evaluation benchmarks still fall short of real agentic complexity. In the Indian setting, the scarcity of mature public benchmarks for major Indian languages and limited independent algorithmic auditing capacity constrain both research and oversight. Liability attribution when an autonomous assistant causes harm remains largely untested in Indian courts.

Conclusion

Multi-domain AI assistants concentrate capability and, with it, risk. The literature reviewed in this paper demonstrates that technical vulnerabilities (prompt injection, excessive agency, hallucination, data leakage), socio-technical harms (bias, misinformation, representational injury), and systemic threats (supply-chain compromise, multi-agent cascade failures) are interrelated and amplified by the very features that make these systems useful. India’s rapid adoption, combined with the specific contours of the DPDP Act, linguistic diversity, and digital public infrastructure, produces a distinctive risk environment that existing global research only partially addresses.

Progress will depend on layered defenses that do not place sole reliance on model alignment, on governance mechanisms that can keep pace with agentic capabilities, and on research that centers the needs and constraints of diverse user populations. The comparative review and risk synthesis offered here are intended as a foundation for such continued work. As multi-domain assistants move from experimental tools to infrastructure, the quality of risk management will determine whether their benefits can be realized safely and equitably.

References

[1] Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security. https://arxiv.org/abs/2302.12173

[2] Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., ... & Liu, Y. (2023). Jailbreaking ChatGPT via prompt engineering: An empirical study. arXiv preprint arXiv:2305.13860. https://arxiv.org/abs/2305.13860

[3] Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How does LLM safety training fail? Advances in Neural Information Processing Systems, 36. https://arxiv.org/abs/2307.02483

[4] Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., ... & Irving, G. (2022). Red teaming language models with language models. arXiv preprint arXiv:2202.03286. https://arxiv.org/abs/2202.03286

[5] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., ... & Raffel, C. (2021). Extracting training data from large language models. In 30th USENIX Security Symposium (pp. 2633–2650). https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting

[6] Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). https://dl.acm.org/doi/10.1145/3442188.3445922

[7] Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P. S., Mellor, J., ... & Gabriel, I. (2022). Taxonomy of risks posed by language models. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 214–229). https://dl.acm.org/doi/10.1145/3531146.3533088

[8] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258. https://arxiv.org/abs/2108.07258

[9] OWASP GenAI Security Project. (2025/2026). OWASP Top 10 for Large Language Model Applications. https://genai.owasp.org/

[10] Zou, A., Wang, Z., Kolter, J. Z., & Fredrikson, M. (2023). Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15043. https://arxiv.org/abs/2307.15043

[11] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2023). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv preprint arXiv:2311.05232. https://arxiv.org/abs/2311.05232

[12] Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., ... & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. Computational Linguistics, 50(3), 1097–1179. https://doi.org/10.1162/coli_a_00524

[13] Yao, Y., et al. (2024). A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly. High-Confidence Computing. https://arxiv.org/abs/2312.02003

[14] Deng, Y., et al. (2024). AI agents under threat: A survey of key security challenges and future pathways. ACM Computing Surveys. https://doi.org/10.1145/3716628

[15] Ko, R., Chung, J., Zheng, S., et al. (2026). Seven security challenges in cross-domain multi-agent LLM systems. npj Artificial Intelligence. https://www.nature.com/articles/s44387-026-00128-9

[16] Li, X., Zhao, S., Zou, X., Zhao, F., Liu, F., Zhang, Y., ... & Wang, J. (2026). Evolving safety landscape of multi-modal large language models: A survey of emerging threats and safeguards. arXiv preprint arXiv:2608.07535. https://arxiv.org/abs/2608.07535

[17] Uuk, R., Gutierrez, C. I., Guppy, D., Lauwaert, L., Kasirzadeh, A., Velasco, L., ... & Prunkl, C. (2024). A taxonomy of systemic risks from general-purpose AI. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5030173

[18] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565. https://arxiv.org/abs/1606.06565

[19] Hendrycks, D., Carlini, N., Schulman, J., & Steinhardt, J. (2021). Unsolved problems in ML safety. arXiv preprint arXiv:2109.13916. https://arxiv.org/abs/2109.13916

[20] Shelby, R., Rismani, S., Henne, K., Moon, A., Rostamzadeh, N., Nicholas, P., ... & Rostamzadeh, N. (2023). Sociotechnical harms of algorithmic systems: Scoping a taxonomy for harm reduction. In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. https://doi.org/10.1145/3600211.3604676

[21] Ministry of Electronics and Information Technology (MeitY) & Office of the Principal Scientific Adviser. (2025/2026). India AI Governance Guidelines and Strengthening AI Governance through Techno-Legal Framework [White paper]. Government of India.

[22] Delinea. (2026). Indian enterprises have AI policies, but agents are still overreaching [Industry survey report].