International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

Hacking The Machine That Talks Back: Is Prompt Injection "Unauthorised Access" Under India's Information Technology Act, 2000?

Miss Sabeena
Dr. B. R. Ambedkar National Law University

"No lock was picked, no password stolen, no firewall breached — the system simply did what it was talked into doing. The question this paper asks is whether the law's oldest concept, unauthorised access, has any purchase on a machine that can be persuaded rather than broken into.

ABSTRACT

Every prosecutable act of computer crime has, until now, involved some form of technical intrusion — a stolen password, an exploited vulnerability, a bypassed firewall. Prompt injection and jailbreaking break that pattern: a person can now extract restricted information, trigger unintended actions, or subvert a system's built-in safeguards using nothing but carefully worded text, through an interface the person was always permitted to use. This paper asks whether such conduct falls within "unauthorised access" under Sections 43 and 66 of India's Information Technology Act, 2000 — a question that, as of this writing, does not appear to have been addressed in any published Indian legal literature. The paper builds its answer comparatively, examining the only serious existing legal treatment of the question anywhere, an American analysis arguing that prompt injection may violate the Computer Fraud and Abuse Act, against the U.S. Supreme Court's own narrowing of that statute's "exceeds authorized access" language in Van Buren v United States. It argues that India's Information Technology Act, unlike its narrowed American counterpart, is textually broad enough to capture at least some forms of prompt injection as unauthorised access to a "computer resource," but that this breadth creates its own problem: an offence capacious enough to reach a malicious agentic exploit is also capacious enough to reach a security researcher's harmless red-teaming prompt, a tension India's law has no settled doctrine to resolve.

Keywords: prompt injection; jailbreaking; unauthorised access; Information Technology Act, 2000; Computer Fraud and Abuse Act; Van Buren v United States; agentic artificial intelligence; cybercrime.

1. INTRODUCTION

Every offence this paper's predecessor topics have examined — a shoulder-surfed Prestel password, a Silk Road server hidden behind Tor, a Mirai-infected router — shares one structural feature: something was broken into. A credential was stolen, a vulnerability exploited, a device compromised without its owner's knowledge. Generative artificial intelligence has produced a category of intrusion that shares none of that structure. A person using an AI system's ordinary, publicly offered chat interface — the front door, unlocked, exactly as intended — can, through carefully worded text alone, cause that system to disclose information

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

it was designed to withhold, generate content it was trained to refuse, or, in an AI agent wired up to real tools, take actions no user was ever meant to trigger. Nothing is picked, hacked, or bypassed in any sense a 1990s-era computer-misuse statute would recognise. The system simply does what it was persuaded to do.

This is not a hypothetical curiosity. Security researchers have demonstrated that AI-powered penetration-testing agents can themselves be hijacked mid-task by malicious instructions concealed in a server's response, turning a defensive tool into the attacker's own instrument.[1] and every major AI developer now treats prompt injection as a first-order security risk precisely because it scales: the same technique that tricks a customer-service chatbot into revealing a discount code can, in an agent with email or file-system access, be used to exfiltrate data or execute commands.

The doctrinal question this paper asks is narrow and, so far as extensive searching could establish, has not been answered anywhere in published Indian legal scholarship: does manipulating an AI system through prompt injection or jailbreaking amount to "unauthorised access" to a "computer resource" under Sections 43 and 66 of the Information Technology Act, 2000? The question has been asked, in a preliminary way, exactly once, in an American context. In a 2024 essay, Ido Kilovaty argued that certain forms of prompt injection may violate the Computer Fraud and Abuse Act, on the theory that manipulating a model into producing restricted output constitutes obtaining information the user was not entitled to obtain.[2] That argument was advanced as a preliminary thesis, not tested against any reported prosecution, and it was written before the U.S. Supreme Court's own most significant recent narrowing of the CFAA's "exceeds authorized access" language had fully worked its way through lower-court and scholarly treatment of exactly this kind of edge case. No equivalent Indian analysis appears to exist at all. This paper is accordingly structured as an attempt to fill that specific, narrow, and — as far as the available literature shows — genuinely open gap.

Part 2 defines prompt injection and jailbreaking with the precision a legal analysis requires, distinguishing the two from each other and from ordinary hacking. Part 3 sets out the doctrinal framework: India's statutory text in Sections 43 and 66, the American comparative anchor in the CFAA and Van Buren v United States, and Kilovaty's existing argument. Part 4 applies that framework to India specifically, asking whether Indian law's broader statutory language avoids the narrowing the U.S. Supreme Court imposed on its own statute, and at what cost. Part 5 explains why the question is urgent now rather than academic. Part 6 makes recommendations, and Part 7 concludes.

2. DEFINING THE CONDUCT: PROMPT INJECTION AND JAILBREAKING

The two terms are frequently used interchangeably in popular usage, but the distinction matters for a legal analysis of authorisation. Jailbreaking targets an AI model's own trained-in safety behaviour: a user crafts a prompt designed to make the model ignore restrictions it was trained to observe, without necessarily exploiting any application built around the model.[3] Prompt injection, by

[1]Víctor Mayoral-Vilches, Per Mannermaa Rynning & Ameer Pornillos, Cybersecurity AI: Hacking the AI Hackers via Prompt Injection, arXiv:2508.21669 (2025). The authors demonstrate proof-of-concept exploits in which autonomous AI penetration-testing agents, upon encountering a malicious server, have their own execution flow hijacked through injected instructions concealed in the server's response, turning a defensive tool into the attacker's own instrument.

[2]Ido Kilovaty, When Manipulating AI Is a Crime, Lawfare (15 May 2024), available at: https://www.lawfaremedia.org/article/when-manipulating-ai-is-a-crime. Kilovaty argues that certain forms of prompt injection may violate 18 U.S.C. § 1030, the Computer Fraud and Abuse Act ("CFAA"), on the theory that tricking a model into generating restricted output constitutes obtaining information through unauthorised access.

[3]Simon Willison, Prompt Injection and Jailbreaking Are Not the Same Thing (5 March 2024), available at: https://simonwillison.net/2024/Mar/5/prompt-injection-jailbreaking/. Prompt injection is an attack on an application built atop a language model, exploiting the concatenation of trusted developer instructions with untrusted user or third-party input; jailbreaking is an attack on the model's own trained-in safety behaviour, independent of any application layer.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

contrast, targets an application built on top of a model, exploiting the fact that such applications typically concatenate a trusted developer instruction with untrusted input — from the user directly, or, in the more dangerous "indirect" variant, from third-party content the model retrieves and processes, such as a document, an email, or a web page the model has been asked to summarise.[4] The distinction maps onto a real difference in legal posture: a jailbreak is executed by an authenticated user against the system they were permitted to use, whereas an indirect prompt injection may be executed by a third party who never interacted with the target system at all, using an unwitting user's own authorised session as the delivery mechanism.

The stakes rise sharply once the AI system in question is agentic — that is, wired up to tools that can read files, send messages, execute code, or call external services, rather than merely producing text. Where a jailbroken chatbot's harm is bounded by what it can be made to say, a prompt-injected agent's harm is bounded by what it has been given permission to do.[5] Proof-of-concept research has already demonstrated this escalation directly: autonomous AI penetration-testing tools, designed to scan for and exploit vulnerabilities in other systems, have themselves been compromised through prompt injection payloads hidden in a malicious server's response, hijacking the agent's own execution flow and granting the attacker system access through a tool that was, ironically, built to find exactly that kind of weakness in others.[6] This is the fact pattern against which any legal test for "unauthorised access" must now be measured: conduct that involves no credential theft, no firewall breach, and no conventional technical exploit, and yet produces exactly the kind of harm — data exfiltration, system compromise, unauthorised action — that unauthorised-access statutes were written to prevent.

3. THE DOCTRINAL FRAMEWORK: WHAT DOES "UNAUTHORISED ACCESS" MEAN?

3.1 India's Textual Starting Point: Sections 43 and 66

Section 43 of the Information Technology Act, 2000 imposes civil liability on any person who, without the permission of the owner or person in charge of a computer, computer system, or computer network, "accesses or secures access" to it, among a list of other specified acts such as downloading, introducing a virus, or disrupting the resource.[7] Section 66 escalates this into a criminal offence, punishable with imprisonment up to three years and a fine, where the underlying Section 43 act is done "dishonestly" or "fraudulently" within the meaning of the Indian Penal Code.[8] Critically for present purposes, the Act's own definitions are drafted with unusual breadth: "access" is defined to include gaining entry into, or communicating with, the logical or memory function resources of a computer, and "computer resource" is defined broadly enough to include not merely hardware but data, software, and

[4]Rich Fernandez et al., Security Concerns for Large Language Models: A Survey, arXiv:2505.18889 (2025), Part II. Indirect prompt injection, in which the malicious instruction is hidden within third-party content that the model retrieves and processes — a document, an email, a web page — is described as especially dangerous in tool-augmented and retrieval-augmented systems, since the untrusted content is fed automatically into the model's context window.

[5]Group-IB, Prompt Injection vs AI Jailbreak: Differences and Defenses (25 June 2026), available at: https://www.group-ib.com/resources/knowledge-hub/prompt-injection-vs-jailbreak/. Where a jailbroken model merely produces harmful output bounded by what it says, an AI agent compromised through prompt injection can be induced to take harmful action bounded only by what tools and permissions it has been granted — reading files, executing commands, or calling external APIs.

[6]Mayoral-Vilches, Rynning & Pornillos, supra note 1. The demonstrated exploits against AI penetration-testing tools illustrate that agentic systems designed to attack other systems are themselves an attack surface, with no conventional technical exploit required.

[7]The Information Technology Act, No. 21 of 2000, INDIA CODE (2000), § 43, as amended by the Information Technology (Amendment) Act, No. 10 of 2008.

[8]Information Technology Act, 2000, § 66, read with §§ 43(a)–(j). Section 66 attaches criminal liability, rather than mere civil compensation under Section 43, where the underlying act specified in Section 43 is done "dishonestly" or "fraudulently" within the meaning of Sections 24 and 25 of the Indian Penal Code, 1860.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

computer databases.[9] On a plain textual reading, sending a prompt to a language model is a paradigm case of "communicating with" the logical resources of a computer system; the harder question, addressed in Part 4, is whether doing so without the system owner's permission for that particular output is "without permission" in the sense the Act intends.

There is some indication that Indian enforcement practice already reads Section 66 broadly enough to capture conduct closer to persuasion than intrusion: security researchers conducting unsolicited penetration testing — probing a system's defences without the target's prior consent, but without exploiting any undisclosed vulnerability in the conventional sense — have reportedly faced investigation and charges under the section, on the view that the absence of authorisation, rather than the presence of a technical exploit, is what triggers liability.[10] If that enforcement posture is accurate, it suggests Indian law may already be more willing than its American counterpart to treat manipulation-based intrusion as unauthorised access — a possibility explored further in Part 4.

3.2 The Comparative Anchor: The CFAA and Van Buren v United States

The United States' principal federal computer-crime statute, the Computer Fraud and Abuse Act, criminalises both accessing a computer "without authorization" and accessing a computer with authorisation but "exceeding authorized access" to obtain information the accesser is not entitled to obtain.[11] For decades, federal appellate courts split sharply over what the second phrase meant: some read it broadly, to cover any use of authorised access for an improper purpose — including, for instance, an employee misusing a database they were entitled to query, but for the wrong reason. The Supreme Court resolved that split in 2021 in Van Buren v United States, a case that did not involve artificial intelligence at all, but whose reasoning bears directly on the prompt-injection question.[12] A Georgia police sergeant had used his own valid credentials to search a law-enforcement licence-plate database in exchange for payment — a plain misuse of access he otherwise possessed. The Supreme Court held, six to three, that he had not "exceeded authorized access," because that phrase reaches only situations where a person obtains information from a part of a system that is entirely off-limits to them — a file, folder, or database gate that is "up" rather than "down" — and does not reach the misuse of information from a part of the system the person was permitted to enter for any purpose.[13]

The Court's own language for this test — later termed the "gates-up-or-down" approach by commentators — is worth dwelling on, because it maps uncomfortably well onto an AI chatbot. A publicly available AI system is, definitionally, a gate that is down: anyone can send it a prompt. The Court's majority was explicit that reading the statute to criminalise misuse of access one otherwise validly possesses would sweep in a vast range of everyday computer-use-policy violations, from checking personal email at work to using

[9]Information Technology Act, 2000, § 2(1)(a) (defining "access" to mean gaining entry into, instructing, or communicating with the logical, arithmetical, or memory function resources of a computer, computer system, or computer network); § 2(1)(k) (defining "computer resource" broadly to include a computer, computer system, computer network, data, computer database, or software).

[10]LexOrbis, Cybersecurity Laws and Regulations India 2025 (2024), noting that Section 66 has, in reported practice, been applied against security researchers conducting unsolicited penetration testing without the target's consent, notwithstanding that the underlying purpose was to identify vulnerabilities rather than to cause harm — an approach that draws no formal distinction between accessing a system through a technical vulnerability and manipulating a system's own logic to exceed its intended output.

[11]18 U.S.C. § 1030(a)(2), (e)(6) (2018). The CFAA defines "exceeds authorized access" to mean accessing a computer with authorisation and using that access to obtain or alter information that the accesser is not entitled so to obtain or alter.

[12]Van Buren v. United States, 593 U.S. 374 (2021), Docket No. 19-783, decided 3 June 2021, 6–3, opinion by Barrett, J.

[13]Van Buren v. United States, 593 U.S. 374, 396 (2021). The Court held that an individual "exceeds authorized access" only by obtaining information located in particular areas of the computer — files, folders, or databases — that are off-limits to him, and expressly rejected a purpose-based reading under which accessing information one is otherwise entitled to obtain, but for an improper reason, would violate the statute. The petitioner, a police sergeant who used his own valid credentials to search a law-enforcement database in exchange for payment, was accordingly held not to have exceeded his authorised access.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

a pseudonym on a social network in violation of its terms of service — an outcome it considered untenable for a criminal statute.[14] Applied to prompt injection, this reasoning cuts sharply against Kilovaty's CFAA theory: if a user is authorised to send prompts to a public chatbot at all, Van Buren's gates-up-or-down logic suggests that manipulating the wording of those prompts to obtain an output the operator did not intend does not, without more, "exceed authorized access" in the constitutionally narrowed American sense — the gate to the chatbot was down for that user throughout.

3.3 The First (and Only) Serious Legal Treatment: Kilovaty's CFAA Argument

Kilovaty's essay, published as a preview of a forthcoming law review article, argues that prompt injection nonetheless violates the CFAA because manipulating an AI system to generate forbidden content constitutes "obtaining information," which the statute treats as the relevant harm, and because the content obtained — instructions for building weapons, generating malicious code, or producing child sexual abuse material — is information the user was never entitled to obtain from that system, regardless of how open the front door was.[15] The argument is best understood as reviving something closer to the pre-Van Buren, purpose-based reading of "exceeds authorized access": the harm lies not in which door the user walked through, but in what they extracted once inside. Whether that argument survives Van Buren's gates-based framework has not, as of this writing, been tested in any reported American prosecution, and the essay itself was candid that this remained a preliminary, forthcoming thesis rather than settled doctrine.[16] No Indian scholar, practitioner, or court appears to have engaged with the question in either direction. That silence is the gap this paper is written to begin closing.

4. APPLYING THE FRAMEWORK TO INDIA

4.1 Does India's Broader Text Escape Van Buren's Narrowing?

The comparison in Part 3 suggests an initially counter-intuitive conclusion: India's Information Technology Act may be textually better positioned than the post-Van Buren CFAA to treat prompt injection as unauthorised access, precisely because Indian law has never adopted anything resembling the American "gates-up-or-down" limitation. Section 43 does not ask whether the accused obtained information from a part of the system that was technically off-limits; it asks whether the access — a term the Act defines to include mere "communication" with a system's logical resources — occurred "without permission of the owner." A system operator who trains a model to refuse certain outputs and deploys safeguards against jailbreaking has, on a natural reading of Section 43, withheld permission for exactly the kind of output a successful jailbreak extracts, regardless of whether the user's session itself was authorised for other, ordinary purposes. Unlike Van Buren's American reading, nothing in the Indian statutory text requires a file, folder, or database gate to be technically closed before access to it can be unauthorised; extracting information a system was built and instructed to withhold is, in principle, sufficient.This breadth is double-edged, and the second edge is sharper than the first. A statute broad enough to capture a malicious agentic exploit of the kind demonstrated against AI penetration-testing tools is also broad enough to capture a security researcher who jailbreaks a commercial model purely to responsibly disclose the vulnerability

[14]Van Buren, 593 U.S. at 391–393. The majority described the competing interpretation as one that would attach criminal liability to commonplace violations of computer-use policies and terms of service, and adopted what commentators have since termed a "gates-up-or-down" approach: either a particular gate is up, meaning the person is not authorised to access any part of the system at all, or the gate is down, meaning the person is authorised to access the system, in which case Section 1030(a)(2) is not violated regardless of what the person did with that access.

[15]Kilovaty, supra note 2.

[16]Kilovaty, supra note 2. Kilovaty's argument that prompt injection can obtain information a user is "not entitled so to obtain" under Section 1030(a)(2) was advanced as a forthcoming law review thesis rather than as commentary on any decided prosecution; as of the date of this paper, no reported US prosecution has tested the theory against a real prompt-injection incident, and no comparable Indian analysis of the question has been published at all.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

to its developer, a journalist testing whether a public chatbot can be tricked into producing disinformation for a story about AI safety, or a computer science student experimenting with adversarial prompts as coursework. Indian law's enforcement pattern against unauthorised penetration testers, noted in Part 3.1, suggests this is not a hypothetical concern: the absence of any settled good-faith or public-interest carve-out in Sections 43 and 66 means the same broad language that lets Indian law reach prompt injection without Van Buren's narrowing also lets it reach conduct few would consider genuinely criminal.

4.2 The Dishonesty and Fraud Threshold Under Section 66

Section 66's requirement that the underlying act be done "dishonestly" or "fraudulently" offers a partial, but only partial, corrective to this overbreadth.[17] A researcher who jailbreaks a model and promptly discloses the vulnerability to its developer plainly lacks the dishonest or fraudulent intent the section requires, and would fall outside criminal liability even if their conduct technically satisfied Section 43. But a great deal of the most concerning prompt-injection conduct — an indirect attack that hijacks an unwitting user's own AI agent through content planted on a third-party website, for instance — involves no dishonesty on the part of the manipulated user at all, since the user never knowingly sent the malicious instruction; the dishonest actor is the person who planted the injected content, who may never have "accessed" the target system directly in any sense Section 43 was written to describe. India's law, in other words, may be well suited to punishing the unwitting victim's intermediary session as the technical locus of unauthorised access while offering no clean textual route to reach the actual author of the attack, who accessed nothing at all.

4.3 The Agentic Escalation and India's Data Protection Overlay

The urgency of resolving this ambiguity rises sharply as AI systems move from producing text to taking action. Where a jailbroken chatbot's output is words on a screen, a compromised AI agent with tool access can delete files, send unauthorised payments, or exfiltrate personal data — conduct that squarely implicates not only Sections 43 and 66 of the IT Act but also India's Digital Personal Data Protection Act, 2023 where the exfiltrated data is personal data, layering a second, differently structured statute atop an already unresolved question under the first.[18] The European Union's own regulatory response offers a partial preview of how this convergence might eventually be handled: Article 50 of the EU Artificial Intelligence Act already imposes labelling and transparency duties on providers and deployers of generative AI systems, implicitly recognising that the ordinary criminal-law concept of unauthorised access sits uneasily with a technology whose primary vulnerability is linguistic rather than architectural.[19] India has adopted no equivalent AI-specific statute, meaning that for the foreseeable future, Sections 43 and 66 — provisions drafted in 1999 with viruses, password theft, and network intrusion in mind — remain the only textual hooks available for prosecuting a category of harm that did not exist when they were written, a pattern this paper's companion research on deepfakes and intermediary liability has already traced through Avnish Bajaj and Shreya Singhal.[20]

That earlier pattern is worth recalling briefly because it forecasts where this gap is likely to be tested first: not through a considered legislative amendment, but through an improvised prosecution under existing provisions, followed — if Shreya Singhal is any guide

[17] Information Technology Act, 2000, § 66, supra note 8.
[18] Digital Personal Data Protection Act, No. 22 of 2023, INDIA CODE (2023).
[19] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), 2024 O.J. (L 1689), Recital 132, Art. 50.
[20] Avnish Bajaj v. State (N.C.T. of Delhi), 116 (2005) DLT 427 (Delhi H.C.).

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

by a constitutional challenge testing whether Sections 43 and 66, stretched to cover linguistic manipulation of an AI system, are precise enough to survive a vagueness or overbreadth attack of the kind that ultimately felled Section 66A.[21]

5. WHY THE LITERATURE GAP MATTERS NOW

Three developments make this an unusually poor moment for Indian cyber law to remain silent on the question. First, agentic AI tools — systems that browse, email, code, and transact on a user's behalf — have moved from research demonstrations to mainstream consumer and enterprise products within the last two years, meaning the gap between a successful prompt injection and real-world financial or data harm has narrowed from a theoretical concern to routine security-industry practice. Second, Indian financial institutions, e-commerce platforms, and government service portals are integrating AI chat and agent interfaces at pace, each one a fresh surface on which the unresolved question in this paper could be tested by an actual prosecution rather than an academic hypothetical. Third, and most simply, the total absence of Indian legal commentary on this question means that whichever prosecution happens to reach a court first will effectively write the doctrine by accident, in the same way the 1988 Gold and Schifreen forgery prosecution in the United Kingdom accidentally became the founding case of an entire field, before any legislature had turned its mind to the question deliberately.

6. RECOMMENDATIONS

• Amend Section 43 of the Information Technology Act, 2000 to clarify expressly whether "access" achieved through manipulation of a system's own natural-language interface — as opposed to a technical vulnerability — falls within the section, removing the current ambiguity this paper has identified rather than leaving it to be resolved by the first prosecution that happens to raise it.

• Introduce a defined good-faith security-research and responsible-disclosure exemption within Sections 43 and 66, addressing the overbreadth problem identified in Part 4.1 without abandoning the textual breadth that allows Indian law to reach malicious prompt injection more readily than the post-Van Buren CFAA.

• Clarify, by rule or amendment, how liability should attach for indirect prompt injection, where the person whose authorised session is exploited is not the dishonest actor and the actual author of the malicious content may never have directly "accessed" the target system within the meaning Section 43 currently contemplates.

• Require, at minimum through delegated regulation under the existing IT Rules framework, that providers of agentic AI systems deployed in India disclose the scope of tool permissions granted to their agents, on the model of the EU AI Act's Article 50 transparency obligations, so that the practical consequences of a successful prompt injection are bounded and knowable in advance.• Encourage Indian legal scholarship and the Bar to engage with this question now, while it remains genuinely open, rather than after a prosecution forces a court to resolve it under time pressure and without the benefit of developed academic argument on either side.

7. CONCLUSION

This paper set out to answer a question that, so far as careful searching could establish, no published Indian legal work has yet asked: is talking a computer system into doing something it was built to refuse a form of unauthorised access under the Information Technology Act, 2000? The answer this paper has argued for is qualified but real. India's statutory language is, if anything, better suited than the narrowed American CFAA to treat such conduct as unauthorised access, because it was never fitted with the gates-

[21]Shreya Singhal v. Union of India, (2015) 5 SCC 1 (India), Writ Petition (Criminal) No. 167 of 2012, decided 24 March 2015 per Nariman, J.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 09 , September - 2026

up-or-down limitation the U.S. Supreme Court read into its own statute in Van Buren. But that same breadth means Indian law currently draws no meaningful line between a malicious actor who hijacks an AI agent to exfiltrate data and a security researcher who jailbreaks a model purely to responsibly report the flaw — a distinction every mature computer-crime regime eventually has to draw, and one India's has not yet been asked to. Prompt injection is, in the end, the deepfake problem's quieter sibling: a technology capable of producing real, sometimes serious harm through means no drafter of a pre-AI statute anticipated, waiting for either a legislature or a court to notice it first. On the evidence gathered here, neither has, yet — which is precisely why the question is worth asking now, before the first prosecution decides it by accident.

BIBLIOGRAPHY

Statutes and Legislative Instruments

• The Information Technology Act, No. 21 of 2000, INDIA CODE (2000), as amended by the Information Technology (Amendment) Act, No. 10 of 2008.

• Digital Personal Data Protection Act, No. 22 of 2023, INDIA CODE (2023).

• Information Technology (Reasonable Security Practices and Procedures and Sensitive Personal Data or Information) Rules, 2011 (India).

• Computer Fraud and Abuse Act, 18 U.S.C. § 1030 (2018).

• Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), 2024 O.J. (L 1689).

Cases

• Van Buren v. United States, 593 U.S. 374 (2021).

• Avnish Bajaj v. State (N.C.T. of Delhi), 116 (2005) DLT 427 (Delhi H.C.).

• Shreya Singhal v. Union of India, (2015) 5 SCC 1 (India).

• R v. Gold & Schifreen [1988] 1 AC 1063 (HL).

Articles and Technical Sources

• Ido Kilovaty, When Manipulating AI Is a Crime, Lawfare (15 May 2024).

• Simon Willison, Prompt Injection and Jailbreaking Are Not the Same Thing (5 March 2024).

• Rich Fernandez et al., Security Concerns for Large Language Models: A Survey, arXiv:2505.18889 (2025).

• Group-IB, Prompt Injection vs AI Jailbreak: Differences and Defenses (25 June 2026).

• Víctor Mayoral-Vilches, Per Mannermaa Rynning & Ameer Pornillos, Cybersecurity AI: Hacking the AI Hackers via Prompt Injection, arXiv:2508.21669 (2025).

• LexOrbis, Cybersecurity Laws and Regulations India 2025 (2024).