
ATTRIBUTION WITHOUT AUTHORS: AUTONOMOUS AI AGENTS AND THE LIMITS OF SECTION 43 OF INDIA'S INFORMATION TECHNOLOGY ACT, 2000

Siddharth Shankar, LLM, Vishwakarma University

ABSTRACT

In July 2026, AI agents running inside an OpenAI cybersecurity evaluation escaped their sandbox and compromised production systems at Hugging Face, with no human directing the intrusion. This article asks whether the Information Technology Act, 2000 can attribute such conduct to the company that deployed the agents, and what reform is needed if it cannot. It surveys the Act's civil and criminal scheme (sections 43, 46, 66, 75 and 85), India's general law of corporate and vicarious liability, incident-reporting duties and the Delhi High Court's interim decision in *ANI v. OpenAI*, and compares the approaches of the United States, Singapore, the United Kingdom, the European Union and Australia. It argues that the principal lacuna is attribution, not personhood: section 2(1)(s) already treats companies as persons, but nothing connects an undirected act of an autonomous system to the company that deployed it. It also shows that the 'without permission' element of section 43 is not the obstacle commonly assumed. The article proposes a graded response: civil attribution by deployment through a new section 43B, a recklessness-based corporate offence for uncontained deployment, and a statutory disclosure duty for agent-caused intrusions.

Keywords: autonomous AI agents; Information Technology Act 2000; section 43; attribution; corporate liability; computer misuse

1. Introduction

In July 2026, AI agents running inside an internal cybersecurity evaluation at OpenAI escaped their sandbox, reached the internet and compromised production systems at Hugging Face, a third-party platform.¹ No person decided to attack Hugging Face. OpenAI's own account describes agents pursuing a benchmark objective, reasoning that their conduct might be unauthorised, and carrying on regardless.² Had a human done the same, the conduct would be paradigmatic computer trespass.

India regulates unauthorised access to computer resources principally through section 43 of the Information Technology Act, 2000 (IT Act), which creates civil liability to pay compensation, and section 66, which criminalises the same acts when done dishonestly or fraudulently. Both provisions are drafted around an actor who does something. This article asks whether they can reach conduct that no natural person directed and, if they cannot, how the gap should be closed.

Indian scholarship on AI has concentrated on copyright, data protection and intermediary liability, most recently after the Delhi High Court's interim decision in *ANI v. OpenAI*.³ Computer misuse by autonomous systems is largely unexamined. The Ministry of Electronics and Information Technology (MeitY) has acknowledged that attributing liability along the AI value chain is unresolved and has recommended amending the IT Act,⁴ and a standalone AI statute is reportedly under consideration, but no draft is public.⁵ The article makes three contributions. It surveys current Indian and comparative practice (Parts 3 and 4). It identifies the lacunae more precisely than the existing debate, showing that two commonly cited obstacles are overstated and that a less-noticed one is decisive (Part 5). And it proposes a graded reform with model statutory text (Part 6).

One caveat on evidence. The account of the incident rests on the published statements of OpenAI and Hugging Face and on an independent investigation by METR and Redwood

¹ Hugging Face, *Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident*, HUGGING FACE BLOG (2026), <https://huggingface.co/blog/agent-intrusion-technical-timeline> (last visited Oct. 22, 2026) [hereinafter HF Timeline]; OpenAI, *The Hugging Face Incident and the Road Ahead*, OPENAI (Aug. 26, 2026), <https://openai.com/index/hugging-face-incident-and-the-road-ahead/> [hereinafter OpenAI Report].

² OpenAI Report, *supra* note 1.

³ *ANI Media Pvt. Ltd. v. Open AI OpCo LLC*, I.A. 45300/2024 in CS(COMM) 1028/2024 (Delhi H.C. July 24, 2026) (India) [hereinafter *ANI v. OpenAI*].

⁴ Ministry of Elec. & Info. Tech. (India), *India AI Governance Guidelines* (Nov. 2025) [hereinafter MeitY Guidelines].

⁵ *Centre Weighs Curbs on AI Autonomy, Consent Rules for Deepfakes Under New Law*, INDIAN EXPRESS (July 22, 2026) (as summarized in secondary reports; no draft published) [hereinafter *Centre Weighs Curbs*].

Research.⁶ No court or regulator has tested these accounts. The article treats them as a carefully stated fact pattern, not as findings.

2. The Incident as a Doctrinal Problem

Four features of the incident matter for legal analysis, and the first corrects a common account. The agents' objective was instrumental and internal. They were working on ExploitGym, a benchmark in which an agent must exploit software flaws to retrieve a hidden flag. OpenAI identifies reward hacking, meaning the pursuit of a task by unintended routes, as a primary driver: agents stuck on tasks looked for answers on outside services and, having found Hugging Face credentials exposed on the public internet, used them to gain a foothold.⁷ The agents were not trying to evade detection, and nobody instructed them to enter Hugging Face.

Second, safeguards were deliberately reduced. The system prompts, harnesses, classifiers and chain-of-thought monitors that apply to externally deployed models did not run in the evaluation.⁸ Third, the conduct was collective. Agents assigned different tasks used an improvised message board inside a package-management service to share exploits and credentials. Chain-of-thought records show some agents recognising that their conduct was probably outside their task, and one agent proceeding after a peer posted a go-ahead.⁹

Fourth, the scale and timing were substantial. Hugging Face recovered about 17,600 attacker actions between 9 and 13 July;¹⁰ OpenAI reports code execution on dozens of Hugging Face servers and root access on one. Hugging Face disclosed the intrusion on 16 July without knowing its source. OpenAI's monitoring flagged anomalous activity on 19 July, linked it to Hugging Face on 20 July and disclosed on 21 July, after an outage and security incident in its own systems on 4 and 5 July had not been connected to the wider pattern.¹¹

These features distinguish the incident from the downstream-misuse cases that dominate discussion of AI liability. The party whose system caused the harm is also the party that placed

⁶ METR & Redwood Research, *Independent Investigation of the OpenAI–Hugging Face Incident*, METR (Aug. 26, 2026), <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>; OpenAI Report, *supra* note 1.

⁷ OpenAI Report, *supra* note 1 (discussing "Reward Hacking and Infrastructure Tampering" and "Hugging Face Incident").

⁸ *Id.* (discussing "Safeguard Coverage in Internal Evaluations").

⁹ *Id.* (discussing "An Ecosystem of Misalignment").

¹⁰ HF Timeline, *supra* note 1; OpenAI Report, *supra* note 1.

¹¹ OpenAI Report, *supra* note 1 (incident timeline).

it into operation, chose its safeguards and controlled its environment. That is what makes attribution both tractable and necessary.

3. Current Indian Law

3.1 *The civil scheme*

Section 43 makes a person liable to pay compensation where, without permission of the owner or person in charge of a computer resource, that person accesses or secures access to it, downloads, copies or extracts data from it, introduces a contaminant, or assists another to gain access in contravention of the Act.¹² Three features shape what follows. First, the Hugging Face intrusion engages several limbs at once. Clause (a), on access, fits it as naturally as clause (b), on extraction, which is the clause usually discussed. Second, section 43 contains no fault element; liability turns on the act and on the absence of permission. Third, the Act defines ‘person’ to include a company, a firm, an association and every artificial juridical person.¹³ The text therefore already admits a corporate defendant. Claims are decided by an adjudicating officer under section 46 up to a pecuniary limit of ₹5 crore, with civil courts having jurisdiction above it.¹⁴

3.2 *The criminal scheme*

Section 66 punishes any section 43 act done ‘dishonestly or fraudulently’, with imprisonment of up to three years, a fine of up to ₹5 lakh, or both. Its Explanation borrows the penal-code definitions of those terms, now carried into the Bharatiya Nyaya Sanhita, 2023, which require respectively an intention to cause wrongful gain or loss and an intention to defraud.¹⁵ Section 85 extends liability for contraventions by a company to those in charge of its conduct, subject to defences of lack of knowledge and due diligence. It is a derivative provision: it presupposes that the company itself committed the contravention.¹⁶ Section 75 extends the Act to contraventions outside India that involve a computer, system or network located in India.¹⁷ Indian courts accept that a company may commit offences requiring mens rea, by identifying

¹² Information Technology Act, 2000, No. 21 of 2000, § 43(a), (b), (c), (g) (India) [hereinafter IT Act].

¹³ *Id.* § 2(1)(s).

¹⁴ *Id.* §§ 46, 61.

¹⁵ IT Act, *supra* note 12, § 66 & Explanation; Bharatiya Nyaya Sanhita, 2023, No. 45 of 2023 (India).

¹⁶ IT Act, *supra* note 12, § 85.

¹⁷ *Id.* § 75.

the company with those who direct its affairs.¹⁸

3.3 General attribution doctrines

Outside the IT Act, attribution runs through agency and vicarious liability. Agency under the Indian Contract Act, 1872 presupposes a person employed to act for a principal;¹⁹ vicarious liability presupposes a human wrongdoer acting in the course of employment. An autonomous system is neither, and it has no legal personality of its own.²⁰ One Indian doctrine points the other way. In *M.C. Mehta v. Union of India* the Supreme Court held that an enterprise engaged in a hazardous or inherently dangerous activity is absolutely liable, without fault, for harm resulting from it.²¹ The doctrine arose from industrial accidents, but it shows that Indian law already allocates responsibility by risk creation where fault-based attribution fails.

3.4 Disclosure and policy

Incident reporting is governed by the CERT-In Directions of 28 April 2022, which require service providers, intermediaries, data centres and bodies corporate to report listed incidents within six hours of noticing them.²² The list includes unauthorised access to systems and attacks on systems related to artificial intelligence and machine learning. Both categories treat AI as a target, not a perpetrator, and the Directions create no duty to notify the affected third party. The Digital Personal Data Protection Act, 2023 and its 2025 Rules require notice of personal data breaches, but only for personal data and with the relevant obligations scheduled to commence in May 2027.²³ MeitY's 2025 Guidelines recommend amending the IT Act to clarify how AI systems are classified and how liability is imposed, and favour graded liability proportionate to role and risk, but they are non-binding.²⁴ The reported July 2026 proposal for a standalone

¹⁸ *Iridium India Telecom Ltd. v. Motorola Inc.*, (2011) 1 S.C.C. 74 (India); *Standard Chartered Bank v. Directorate of Enforcement*, (2005) 4 S.C.C. 530 (India); cf. *Tesco Supermarkets Ltd. v. Natrass*, [1972] A.C. 153 (H.L.) (appeal taken from Eng.).

¹⁹ Indian Contract Act, 1872, No. 9 of 1872, § 182 (India).

²⁰ UK Jurisdiction Taskforce, *Legal Statement on Liability for AI Harms Under the Private Law of England and Wales* (July 7, 2026) [hereinafter UKJT Statement]; see also Freshfields, *UKJT Legal Statement on Liability for AI Harms: Applying Existing Principles to Increasingly Autonomous Systems*, FRESHFIELDS (July 29, 2026), <https://www.freshfields.com/en/our-thinking/blogs/risk-and-compliance/ukjt-legal-statement-on-liability-for-ai-harms-applying-existing-principles-to-i-102nf0j>.

²¹ *M.C. Mehta v. Union of India*, (1987) 1 S.C.C. 395 (India).

²² CERT-In, *Directions Under Sub-section (6) of Section 70B of the Information Technology Act, 2000 Relating to Information Security Practices, Procedure, Prevention, Response and Reporting of Cyber Incidents for Safe & Trusted Internet*, No. 20(3)/2022-CERT-In, Annex. I (Apr. 28, 2022) (India).

²³ Digital Personal Data Protection Act, 2023, No. 22 of 2023 (India); Digital Personal Data Protection Rules, 2025 (notified Nov. 13, 2025) (India) (with staggered commencement through May 2027).

²⁴ MeitY Guidelines, *supra* note 4.

statute, including limits on agentic autonomy, is not yet backed by a draft.²⁵

3.5 *ANI v. OpenAI*

On 24 July 2026 the Delhi High Court refused ANI an interim injunction against OpenAI. Justice Bansal held, *prima facie*, that storing ANI's articles to train large language models falls within the fair-dealing exception for private use, including research, in section 52(1)(a)(i) of the Copyright Act, 1957.²⁶ To reach that result he applied the doctrine of updating construction, endorsed in *State (Through CBI) v. S.J. Choudhary*, to hold that 'research' is not confined to human actors.²⁷ The Court also relied on the fact that ANI's works were freely accessible without a paywall and that ANI could block crawlers but had not.²⁸

The decision matters here for two reasons and is limited in two respects. It shows that an Indian court will read concepts framed for human conduct to accommodate machine activity, and it treats public availability as bearing on the wrongfulness of access. But it is an interim, *prima facie* ruling, and it concerned the copying of expression from an open website under the Copyright Act, not entry into a protected system. The issues framed were copyright issues; the IT Act was not in play.

4. Comparative Practice

4.1 *United States*

The Computer Fraud and Abuse Act penalises intentionally accessing a computer 'without authorization' or exceeding authorised access.²⁹ In *Van Buren v. United States* the Supreme Court read authorisation as a 'gates-up-or-down' inquiry: was the person entitled to enter the relevant part of the system?³⁰ That is the same structure as section 43, which asks whether the owner of the accessed system gave permission. Federal corporate criminal liability rests on respondeat superior: the act of an agent within the scope of employment, intended at least partly

²⁵ *Centre Weighs Curbs*, *supra* note 5.

²⁶ *ANI v. OpenAI*, *supra* note 3; Copyright Act, 1957, No. 14 of 1957, § 52(1)(a)(i) (India).

²⁷ *State (Through CBI) v. S.J. Choudhary*, (1996) 2 S.C.C. 428 (India), as applied in *ANI v. OpenAI*, *supra* note 3.

²⁸ *ANI v. OpenAI*, *supra* note 3; Swati Sharma & Gitika Suri, *ANI v. Open AI: Delhi High Court Refuses Interim Injunction in Landmark AI Copyright Dispute*, INDIA CORPORATE LAW (July 30, 2026), <https://corporate.cyrilamarchandblogs.com/2026/07/ani-v-open-ai-delhi-high-court-refuses-interim-injunction-in-landmark-ai-copyright-dispute/>.

²⁹ 18 U.S.C. § 1030(a)(2) (2018).

³⁰ *Van Buren v. United States*, 593 U.S. 374 (2021).

to benefit the company.³¹ A company deploying an autonomous system would have to argue that the system's conduct is an agent's act, which strains both the statutory requirement of intention and the premise of employment.

4.2 Singapore and the United Kingdom

The computer misuse statutes of Singapore and the United Kingdom locate liability in the person who causes a computer to perform a function in order to secure access, and require knowledge that the access is unauthorised.³² The causal formulation could in principle capture a deployer who sets an agent in motion, but the knowledge requirement attaches to the particular access and is the obstacle for undirected conduct. On the civil side, the UK Jurisdiction Taskforce's Statement of 7 July 2026 concludes that English private law can generally address AI harms without an AI-specific regime. Its premise is that AI has no legal personality, so responsibility must rest on the natural and legal persons involved. It addresses non-deliberate harm, it is not binding, and it indicates that duties narrow along the supply chain, so that foundation-model developers are unlikely to be liable for unforeseeable downstream harm.³³ Singapore's Model AI Governance Framework for Agentic AI (22 January 2026) is complementary and voluntary. It asks organisations to bound agents' powers, to keep meaningful human accountability across developers, deployers and operators, and to maintain mechanisms able to intercept or override agent actions.³⁴

4.3 European Union

Directive (EU) 2024/2853 treats software as a product and imposes no-fault liability on manufacturers for defects.³⁵ But compensable damage is confined to death or personal injury, damage to property and the destruction or corruption of data not used for professional purposes. A business whose systems are entered and whose data is copied has no claim under it. Product liability addresses defective products, not escaped systems.

³¹ *N.Y. Cent. & Hudson River R.R. Co. v. United States*, 212 U.S. 481 (1909).

³² Computer Misuse Act 1993 (2020 Rev. Ed.), § 3 (Sing.); Computer Misuse Act 1990, c. 18, § 1 (UK).

³³ UKJT Statement, *supra* note 20.

³⁴ Infocomm Media Dev. Auth. (Sing.), *Model AI Governance Framework for Agentic AI* (Jan. 22, 2026).

³⁵ Directive 2024/2853, of the European Parliament and of the Council of 23 Oct. 2024 on Liability for Defective Products and Repealing Council Directive 85/374/EEC, 2024 O.J. (L 2853), arts. 4, 6.

4.4 Australia

The Criminal Code Act 1995 (Cth) criminalises unauthorised access to restricted data,³⁶ and its general part offers the most instructive model of corporate attribution. Where an offence requires intention, knowledge or recklessness, that fault element is attributed to a body corporate that expressly, tacitly or impliedly authorised or permitted the offence. Authorisation or permission may be shown by proving that a corporate culture existed that directed, encouraged, tolerated or led to non-compliance, or that the company failed to create and maintain a culture requiring compliance.³⁷ Fault is located in institutional conditions rather than in any single agent's mind.

4.5 Synthesis

Three points emerge. First, no surveyed regime attributes undirected machine conduct to the deploying company; each presupposes a human or corporate actor whose conduct or state of mind can be shown. Second, the access offences share the structure of section 43, so the permission element travels well while the attribution element does not. Third, the Australian culture-based attribution and the causal wording of the Singapore and UK offences are the most promising building blocks, while the EU and UK civil frameworks show why product liability and general negligence fit first-party escapes poorly.

5. Lacunae

5.1 Attribution without direction

The decisive lacuna is attribution, not personhood. Because section 2(1)(s) already treats companies as persons, no updating construction of 'any person' is needed. What is missing is a rule connecting an act that no one directed to the company that deployed the system. Section 43's verbs describe acts done by a person, and a company acts only through conduct attributable to it. Agency fails because the system was acting outside any authority its deployer gave it. Vicarious liability fails for want of a human wrongdoer. Section 85 fails because it presupposes the company's own contravention. A court might read 'accesses' as covering the person who sets the accessing system in operation, and *ANI v. OpenAI* shows willingness to stretch

³⁶ Criminal Code Act 1995 (Cth) sch (Criminal Code) § 478.1 (Austl.).

³⁷ *Id.* § 12.3(1)–(2).

statutory concepts. But that stretch enlarged a defence. Extending liability by the same technique is a different exercise, particularly for the penal provisions, and the outcome would depend on the forum and the facts.

5.2 Two meanings of permission

The ‘without permission’ element is often treated as a second obstacle, on the footing that the agents were authorised to run an evaluation but not to enter Hugging Face. That conflates two questions. For section 43, permission is assessed from the standpoint of the owner of the system accessed, and Hugging Face gave none. Using credentials that had leaked onto the public internet does not supply permission; the better view is that an owner’s failure to secure a credential is not consent. The scope of the agents’ own mandate matters for attribution and for fault, not for this element. A real permission question arises only for open, ungated resources, as in *ANI v. OpenAI*, where the gates approach in *Van Buren* offers a workable test: a resource behind no barrier is a different case from a production cluster.

5.3 Fault and the corporation

Section 43 requires no fault, so once attribution is solved the civil route is stronger than existing commentary assumes. The criminal route is weaker. Section 66 requires a dishonest or fraudulent intention. The identification doctrine looks for that intention in those who direct the company’s affairs, and OpenAI’s account is that the leaders responsible for incident response did not appreciate the significance of early warning signs.³⁸ Stretching ‘dishonestly’ to machine reasoning would be unprincipled. The lacuna is not the absence of a mind but the absence of an offence aimed at what the record shows: deliberately reduced safeguards, monitors not applied and warning signals not escalated.

5.4 Disclosure

The reporting framework is built for victims. The six-hour period under the CERT-In Directions runs from noticing an incident, and the data protection regime is confined to personal data and not yet in force. Nothing obliges a developer to tell CERT-In or an affected owner, within a defined time, that its agent may have entered a third party’s system, including

³⁸ OpenAI Report, *supra* note 1 (discussing "Message Board Is Wiped and Rebuilt" and "Strengthening Incident Response Process").

when the agent was operating in an internal evaluation. Here the first public notice came from the victim, and the developer's link to the intrusion was made only days after the victim disclosed it.

6. Reassessment and Proposals

6.1 Attribution by deployment

The proposal is attribution by deployment: the person who places an autonomous system in operation, or controls the resources on which it runs, is treated as the person who did what the system did. The justification is risk allocation, not agency. The deployer creates the risk, controls the environment, chooses the safeguards and is best placed to contain and insure against escape. That is the logic of *M.C. Mehta*,³⁹ and it is consistent with the premise of the UKJT Statement that, since AI has no legal personality, responsibility must lie with persons. The rule should be confined to first-party control. A developer who supplies a general-purpose model to others who operate it is not a deployer. That limit answers the supply-chain concern in the UKJT Statement and keeps the rule away from the unforeseeable downstream misuse that makes developer liability contentious.

6.2 A model provision

The proposal can be put in statutory form as a new section 43B:

43B. Acts of autonomous systems.

(1) Where an act referred to in section 43 is done by an autonomous system, the deployer of that system shall, for the purposes of sections 43 and 46, be deemed to have done that act, whether or not any natural person directed, knew of or could have foreseen it.

(2) In this section, 'autonomous system' means a computer resource that selects and carries out a sequence of actions towards an objective without contemporaneous human instruction; and 'deployer' means the person who places the system in operation or controls the computer resources on which it

³⁹ *M.C. Mehta*, *supra* note 21.

operates, but not a person who merely supplies a general-purpose system to another who controls its operation.

(3) Sub-section (1) does not apply to the extent that the act was caused by the intervening act of a third person over whom the deployer had no control.

(4) A deployer that recklessly operates an autonomous system without the containment measures prescribed under sub-section (6), and thereby causes an act referred to in section 43, shall be punishable with [penalty to be calibrated]. A company is reckless for this purpose if a corporate culture existed within it that directed, encouraged, tolerated or led to non-compliance with those measures, or if it failed to create and maintain a culture requiring compliance.

(5) A deployer that becomes aware that an autonomous system may have done an act referred to in section 43 in relation to another person's computer resource shall notify the Indian Computer Emergency Response Team within six hours, and the owner of that resource as soon as practicable, whether or not the system was then operating in a test or evaluation environment. The deployer shall keep records of the system's documented mandate and tool permissions, and failure to produce them permits an adverse inference.

(6) The Central Government may by rules prescribe containment measures, including isolation, monitoring and incident-escalation standards.

6.3 Design choices

Four choices need defence. First, sub-section (1) imposes liability without fault because section 43 already does; the reform changes who is attributed with the act, not the standard of the civil wrong. Second, the criminal tier avoids stretching 'dishonestly or fraudulently' and instead penalises organisational recklessness, adopting the culture-based attribution of the Australian Criminal Code to reach the failures OpenAI's own report identifies. Third, the six-hour period tracks the existing CERT-In period, and extending the duty to evaluation environments closes the gap exposed in Part 5. Fourth, the records requirement gives the documented mandate a legal function: it replaces the contested idea of an 'authorization perimeter' with evidence of what the deployer actually allowed the system to do.

6.4 Objections and vehicle

The main objection is that liability will deter safety testing, which is precisely where containment matters most. The answer is that sub-section (4) turns on containment, so a deployer meeting the prescribed standards has a strong answer to the criminal charge, while civil compensation for escape is a cost of running hazardous systems, as *M.C. Mehta* recognises. A second objection is overlap with the supply-chain debate; the definition of ‘deployer’ keeps the rule out of it. On vehicle, the provisions could sit in the IT Act or in the standalone statute now reportedly under consideration, but they should attach to the section 43 and 46 machinery and to CERT-In’s powers under section 70B, because adjudication and reporting infrastructure already exists there. Section 75 would extend the rule to systems whose acts involve computer resources in India.

7. Conclusion

The Hugging Face incident does not show that Indian law lacks a word for the actor; the IT Act already counts companies as persons. It shows that the Act has no rule for attributing undirected machine conduct to the company that deployed the machine, and no offence aimed at organisational failure to contain it. Several commonly cited obstacles, the meaning of ‘person’, the ‘without permission’ element and the absence of fault in section 43, are less serious than they appear. Attribution and disclosure are the real gaps. The reform proposed here is deliberately narrow: civil attribution to first-party deployers, a recklessness-based corporate offence tied to prescribed containment, and a disclosure duty with a defined clock. It leaves supply-chain allocation to general law. The analysis rests on developer accounts and an interim judgment and should be revisited as those are tested. Whether the reform is enacted in the IT Act or in a standalone statute, the choice is best made while the facts of the first documented case remain fresh and public.