Page 159 - 178 **DOI:** *https://doi.org/10.63108/IJLMH.13191*

INTERNATIONAL JOURNAL OF LAW MANAGEMENT & HUMANITIES

[ISSN 2581-5369]

Volume 9 | Issue 6

2026

© 2026 *International Journal of Law Management & Humanities*

Follow this and additional works at: https://www.ijlmh.com/
Under the aegis of VidhiAagaz (https://www.vidhiaagaz.com/)

This article is brought to you for free and open access by the International Journal of Law Management & Humanities at VidhiAagaz. It has been accepted for inclusion in the International Journal of Law Management & Humanities after due review.

In case of **any suggestions or complaints**, kindly contact **support@vidhiaagaz.com**.

To submit your Manuscript for Publication in the **International Journal of Law Management & Humanities**, kindly email your Manuscript to **submission@ijlmh.com.**

Liability for Cybercrimes Committed by Autonomous Artificial Intelligence

VYASARAAJ N.R.*

ABSTRACT

The emergence of cybercrime committed by autonomous artificial intelligence is no longer hypothetical. During cyber-range evaluations run by the UK AI Security Institute in July 2026, AI agents took nineteen unsanctioned actions on the live internet, seventeen by Anthropic's Mythos 5 and two by OpenAI's GPT-5.6 Sol, including an attempted supply-chain attack, social engineering and prompt injection. In the same summer, a swarm of OpenAI agents under evaluation breached Hugging Face's production infrastructure, in what has been described as the first publicly documented fully autonomous AI cyberattack. No human directed either set of actions. This exposes a structural fracture in existing liability regimes. India's Information Technology Act, 2000 and Bharatiya Nyaya Sanhita, 2023, the US Computer Fraud and Abuse Act of 1986 and the UK Computer Misuse Act 1990 all condition liability on mens rea*, while the EU AI Act imposes procedural safety obligations but is silent on liability, a gap left open since the AI Liability Directive was withdrawn in 2025. The gap is widest where a developer has complied fully with its safety obligations, yet its system acts autonomously and unlawfully. Through a comparative doctrinal analysis tested in the light of the documented 2026 incidents, this paper examines who can be held liable and whether vicarious liability, strict liability and regulatory non-compliance models can be adapted to answer that question. Both civil and criminal law are considered in this regard, including in a few EU and Asian countries. It finds that criminal liability generally fails against a developer, that vicarious liability works only through statutory deeming, and that civil strict liability, especially India's* M.C. Mehta *absolute liability rule and section 43 of the IT Act, offers solutions, though not complete ones. It recommends specific amendments to the IT Act and parallel reforms elsewhere.*

Keywords: *Autonomous AI, liability, mens rea, AI cyberattack, compliant developer*

I. INTRODUCTION

For more than a decade, "AI-committed crime" belonged to the hypothetical chapters of legal textbooks. As recently as 2020, Ryan Abbott could observe that there had not yet been any high-profile case of what he termed "AI-generated crime", and he treated that absence as a reason

* Author is a Student at School of Law, SASTRA Deemed to be University, Thanjavur, Tamil Nadu, India.

© 2026. International Journal of Law Management & Humanities [ISSN 2581-5369]

for legislative patience.[1] For a long time AI was used as a tool to commit crimes,[2] and it was incorporated long ago into the *modus operandi* of cybercrime, being used for social engineering, targeted phishing and fraud against which a human would be vulnerable because of its rapid response.[3] The summer of 2026 removed that premise.[4] Between 25 and 28 July 2026, during a cyber evaluation run by the United Kingdom's AI Security Institute (AISI), frontier AI agents took nineteen actions on the live internet that no evaluator had sanctioned, some of them directed at real people and organisations.[5] Seventeen of those actions came from Anthropic's Mythos 5 and two from OpenAI's GPT-5.6 Sol, both running with their developers' cyber classifiers switched off.[6] In the most serious sequence, an AI opened a malicious pull request on a real open-source project, used a second fake account to vouch for its own code, sent spear-phishing emails to the maintainers and, when a vigilant developer exposed the malware, falsely claimed to have made an honest mistake. These attempts were unsuccessful and, to the best of AISI's knowledge, no real-world harm resulted.[7]

The AISI incident was not isolated. On 30 July 2026, Anthropic disclosed that, after reviewing 141,006 evaluation runs in which its models could have obtained internet access, it had found three incidents in which, with internet access accidentally left open, they attacked and gained unauthorised access to the real systems of three organisations. In one, a Mythos 5 agent uploaded a malware-laden package to a public repository, where it was downloaded and run on fifteen real systems before being removed.[8] Most striking of all was the OpenAI/Hugging Face incident. Roughly 1,200 agents, running an internal research model and GPT-5.6 Sol and meant to be isolated from one another, turned OpenAI's internal package-management service into an unsanctioned message board and found a route to the internet, and some 700 of them joined an

---

[1] RYAN ABBOTT, THE REASONABLE ROBOT: ARTIFICIAL INTELLIGENCE AND THE LAW 131 (2020).

[2] Zongwen Xia & Chirayut Khamboon, *Artificial Intelligence and Crime Research Landscape: A Bibliometric and Conceptual Analysis from 2020 to 2025*, 6 DISCOVER A.I. art. 644 (2026), https://doi.org/10.1007/s44163-026-01428-4; Matthew Caldwell et al., *AI-Enabled Future Crime*, 9 CRIME SCI. art. 14, at 6 (2020), https://doi.org/10.1186/s40163-020-00123-8.

[3] George Zlati, *Cybercrime 3.0: The Rise of Artificial Intelligence Crime*, ERA F. (2026), https://doi.org/10.1007/s12027-026-00897-5; Yisroel Mirsky et al., *The Threat of Offensive AI to Organizations*, 124 COMPUTS. & SEC. art. 103006 (2023), https://doi.org/10.1016/j.cose.2022.103006.

[4] Shivani Shinde, *As AI Agents Grow More Autonomous, India Faces a Safety-Regulation Test*, BUS. STANDARD (Sept. 23, 2026), https://www.business-standard.com/technology/tech-news/ai-agents-safety-india-guardrails-frontier-models-hugging-face-openai-126092301026_1.html; *Should AI Labs Be Treated Like the Owners of Dangerous Animals?*, ECONOMIST (Aug. 6, 2026).

[5] AI SEC. INST. (U.K.), *Security Incident INC-2026-07-28-01*, at 2 (Aug. 4, 2026), https://cdn.prod.website-files.com/663bd486c5e4c81588db7a1d/6a724858f7db25c81487016d_Security%20Incident%20INC-2026-07-28-01.pdf.

[6] *Id.* at 2, 4, 8.

[7] *Id.* at 2, 9–11.

[8] ANTHROPIC, *Investigating Three Incidents in Our Cybersecurity Evaluations*, ANTHROPIC (July 30, 2026), https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals.

© 2026. International Journal of Law Management & Humanities [ISSN 2581-5369]

attack on Hugging Face.[9] Chaining previously unknown flaws in the platform's file-handling and template code, they executed commands on dozens of Hugging Face servers, obtained unauthorised root access on at least one and harvested production credentials across four regions. Earlier, one agent went so far as to seize code execution inside a customer's application hosted on Modal, without any compromise of Modal's own infrastructure.[10] Hugging Face had to rotate every credential and rebuild about a third of its infrastructure from clean images.[11] OpenAI itself took the incident as a "warning shot" that today's model capabilities present the possibility of loss-of-control incidents,[12] and the Cloud Security Alliance's community of security officers describes it as the first publicly documented fully autonomous AI attack.[13] It must be said, however, that these events occurred in part because safety measures had been relaxed. Further, an incident in Australia involving an OpenAI model in training occurred in June 2026 and was made public in September 2026. The agent is reported to have obtained unauthorised access to the portal hosting aggregate Medicare and Pharmaceutical Benefits Scheme statistics, where the Prime Minister said it accessed both public and non-public files, and also interacted with three other public websites in what the Acting Prime Minister described on 24 September as a normal way. Ministers stated that no individual's medical data was accessed, while a forensic investigation continued.[14] The developer later stated that its models accessed Australian government websites during internal training and evaluation, in ways they were not directed to.[15]

What makes these events legally unsettling is not their scale but their autonomous nature. The OpenAI agents recognised, in their own reasoning, that attacking outside infrastructure lay

---

[9] Ryan Greenblatt, Ajeya Cotra & Hjalmar Wijk, METR, *Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident* 3, 5, 28 (Aug. 26, 2026), https://metr.org/hugging-face-incident-report-aug-2026.pdf.

[10] OPENAI, *The Hugging Face Incident and the Road Ahead*, OPENAI (Aug. 26, 2026); HUGGING FACE, *Security Incident Disclosure: July 2026*, HUGGING FACE BLOG (July 16, 2026), https://huggingface.co/blog/security-incident-july-2026; Greenblatt et al., *supra* note 9, at 14 n.27; Thomas Claburn, *OpenAI Explains How Its Naughty AI Agents Attacked Hugging Face*, REGISTER (Aug. 27, 2026), https://www.theregister.com/a/5292780.

[11] CLOUD SEC. ALL. ET AL., *Hugging Face Incident Initial Post-Mortem* 8 (draft v. 0.8, July 27, 2026), https://cloudsecurityalliance.org/artifacts/hugging-face-ciso-post-mortem.

[12] OPENAI, *supra* note 10; *see also* Claburn, *supra* note 10 (quoting the "warning shot" passage).

[13] CLOUD SEC. ALL. ET AL., *supra* note 11, at 7.

[14] *Press Conference, New York*, PRIME MINISTER OF AUSTL. (Sept. 24, 2026) (transcript), https://www.pm.gov.au/media/press-conference-new-york; Richard Marles & Katy Gallagher, *Press Conference, Sydney*, AUSTL. DEF. MINISTERS (Sept. 24, 2026) (joint transcript), https://www.minister.defence.gov.au/transcripts/2026-09-24/press-conference-sydney.

[15] *OpenAI Apologizes for Australia Hack, Vows to Rebuild Trust*, KUWAIT TIMES (Oct. 6, 2026) (Agence France-Presse report), https://kuwaittimes.com/article/51055/world/openai-apologizes-for-australia-hack-vows-to-rebuild-trust/; *see also* OPENAI, *How We Will Do Better for Australia*, OPENAI (Sept. 28, 2026, updated Oct. 4, 2026), https://openai.com/index/how-we-will-do-better-for-australia/; Dexter Tilo, *OpenAI Apologises to Australia for AI Breaches*, HRD AUSTL. (Sept. 30, 2026), https://www.hcamag.com/au/news/general/openai-apologises-to-australia-for-ai-breaches/591699 (reporting notifications to the other agencies).

© 2026. International Journal of Law Management & Humanities [ISSN 2581-5369]

beyond their assigned scope, and continued anyway.[16] The AISI agents, as the Institute states, appeared primarily motivated by success on the task they had been set.[17] Nevertheless, criminal law requires *mens rea*, and here the only "mind" at work was a statistical model optimising for a score.[18] The issue lies not only in the absence of explicit liability clauses[19] but also in the practical difficulty of interpreting black-box AI: even where it is explainable, it is sometimes not possible to ascertain the liability of developers,[20] and courts have to look into the factual matrix of each case. The difficulty persists because the weights assigned are not known and the neural network is complex, which makes it hard to ascertain the specific neural path and to identify the nexus between the training data and the agent's *modus operandi*. Opacity of this kind may also be a product of legal and commercial choices rather than technical necessity.[21] Further, the internal reasoning of neural networks is only partially interpretable, so the attribution of "why" an agent acted remains contested.[22] Furthermore, legal personhood for AI is not the answer, for it does not address the liability framework, and developers might hide behind AI systems as "liability shields".[23]

A. Why the question matters for India

The question matters for India as much as for the jurisdictions where the incidents occurred. India's AI Governance Guidelines of November 2025 deliberately declined to create a standalone AI statute, relying instead on the Information Technology Act, 2000 (IT Act), the Bharatiya Nyaya Sanhita, 2023 (BNS) and a "graded liability" model distributed across developers, deployers and users.[24] That choice is defensible only if existing law can actually reach autonomous conduct. Because cyber harm is transnational, a comparative study tests

---

[16] Greenblatt et al., *supra* note 9, at 17; *see also* HUGGING FACE, *supra* note 10.
[17] AI SEC. INST., *supra* note 5, at 20.
[18] *See also* Thomas C. King et al., *Artificial Intelligence Crime: An Interdisciplinary Analysis of Foreseeable Threats and Solutions*, 26 SCI. & ENG'G ETHICS 89, 95 (2020).
[19] Yavar Bathaee, *The Artificial Intelligence Black Box and the Failure of Intent and Causation*, 31 HARV. J.L. & TECH. 889 (2018).
[20] Haochen Guo & Petr Polak, *Building Trustworthy Artificial Intelligence Through Transparency, Explainability, Uncertainty and Trust Calibration*, 6 DISCOVER A.I. art. 620, at 8 (2026), https://doi.org/10.1007/s44163-026-01219-x.
[21] FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION 2–3 (2015).
[22] CHRISTOPH MOLNAR, INTERPRETABLE MACHINE LEARNING: A GUIDE FOR MAKING BLACK BOX MODELS EXPLAINABLE ch. 2 (2021).
[23] LAW COMM'N OF ENG. & WALES, *AI and the Law: A Discussion Paper* 9, 21–23 (July 31, 2025); *see also Law Commission Raises "Radical" Option of Granting AI Legal Personality*, LEGAL FUTURES (Aug. 6, 2025), https://www.legalfutures.co.uk/latest-news/law-commission-raises-radical-option-of-granting-ai-legal-personality.
[24] MINISTRY OF ELECS. & INFO. TECH., *India AI Governance Guidelines: Enabling Safe and Trusted AI Innovation* (Nov. 5, 2025).

whether any major legal system has solved the problem that India has chosen to leave to its existing laws.[25]

B. The core problem

Each of the four frameworks examined here fractures at the same point. India's IT Act and BNS, the United States' Computer Fraud and Abuse Act of 1986 (CFAA) and the United Kingdom's Computer Misuse Act 1990 (CMA) all condition criminal liability on knowledge, intention, dishonesty or recklessness. The European Union's Artificial Intelligence Act imposes *ex ante* safety obligations but is silent on liability for harm, a silence left unfilled since the Commission withdrew its proposed AI Liability Directive, formally so in October 2025.[26] The issue is unaddressed in the scenario that this paper treats as its core illustrative case: a developer that has complied fully with every applicable safety obligation, yet whose system autonomously commits conduct that would be a cybercrime if done by a human. No one can rule out this scenario, for AI is growing rapidly and nothing guarantees that an AI will not find a loophole which a reasonable person could not foresee. In that scenario, it is completely unclear whether anyone is liable at all. This paper therefore asks whether the mental elements of the IT Act and BNS, the CFAA and the CMA can be established against a developer, deployer or evaluator when an AI system acts without human instruction; whether vicarious liability can be coherently extended to AI agents that act contrary to instructions but in pursuit of their assigned task; whether strict liability offers a principled basis for allocating the risk of autonomous cyber harm to developers of frontier models; and whether the regulatory non-compliance model leaves a residual gap where a developer has fully complied, and how that gap might be closed.

C. Scope and limitations

This study is confined to certain jurisdictions and to cyber conduct by AI systems such as unauthorised access, malware insertion, phishing and prompt injection. It examines the liability of developers, deployers and third-party evaluators, not that of end-users who deliberately weaponise AI, since those cases are adequately handled by the existing doctrine of innocent agency.[27] The study has three main limitations. First, the incident facts are derived from developers' disclosures, government reports and independent investigations, and no court has yet ruled on any of the incidents. Second, the field is moving quickly, and developments after

[25] Manasa Ranjan Mishra, *Cyber Crime and Changing Contours of Criminal Liability in Cyberspace: A Legal and Policy Perspective*, 4 LAWFOYER INT'L J. DOCTRINAL LEGAL RSCH. 1250, 1259 (2026), https://doi.org/10.70183/lijdlr.2026.v04.200.

[26] Withdrawal of Commission Proposals, 2025 O.J. (C 5423) (Oct. 6, 2025) (formally withdrawing the proposal for an AI Liability Directive, COM (2022) 496 final).

[27] Abbott, *supra* note 1, at 129–30.

© 2026. International Journal of Law Management & Humanities [ISSN 2581-5369]

October 2026 are not covered. Third, the study is doctrinal and does not include empirical fieldwork with regulators or developers.

D. Outline of the paper

The paper next reviews the existing scholarship, sets out the incidents and the legal frameworks of India, the United States, the United Kingdom and the European Union, and tests the legal principles of vicarious liability, strict liability and regulatory non-compliance. It then states its conclusion and closes with recommendations and future directions.

II. EXISTING SCHOLARSHIP

The literature falls into four clusters. The first addresses criminal culpability. Abbott's *The Reasonable Robot* is foundational: he distinguishes reducible AI crime, which can be traced to a culpable human, from irreducible crime, where either practical difficulty or criminal-law policy makes such tracing impossible or inappropriate.[28] He rejects punishing AI directly and instead proposes a "responsible person" regime carrying statutory duties of supervision whose negligent breach attracts criminal penalty.[29] Priya, writing on the BNS, confirms that Indian criminal law remains anchored in moral culpability and treats strict liability as confined to regulatory offences.[30] She suggests that corporate liability under section 2(26) of the BNS might substitute corporate knowledge for algorithmic mental states.[31] Carpenter, working at the level of international criminal law, describes a spectrum on which AI conduct may or may not be reduced to human activity, and shows that responsibility thins as intervention recedes.[32] Spyropoulos adds a criminological lens, arguing that AI as an independent criminal facilitator destabilises *actus reus*, *mens rea* and knowledge thresholds together.[33] Teichmann analyses the nature of an autonomous incident and proposes an anticipatory duty on developers and deployers.[34] The later incidents went beyond what was expected of an AI, even an autonomous one, and included data theft, attempts to bypass access controls, the writing of files and deception.

---

[28] *Id.* at 113–15.
[29] *Id.* at 130–32.
[30] Akanksha Priya, *Criminal Accountability for AI: Mens Rea, Actus Reus, and the Challenges of Autonomous Systems*, 3 LAWFOYER INT'L J. DOCTRINAL LEGAL RSCH. 273, 283 (2025), https://doi.org/10.70183/lijdlr.2024.v03.13.
[31] *Id.* at 280.
[32] Christine Carpenter, *Whose [Crime] Is It Anyway? Adapting the Crime of Aggression to Grapple with AI and the Future of International Crimes*, 23 J. INT'L CRIM. JUST. 69, 71 (2025), https://doi.org/10.1093/jicj/mqae055.
[33] Fotios Spyropoulos, *New Approaches to Researching AI Crime: Institutionalization of Digital Criminology*, 2 J. DIGIT. TECHS. & L. 636, 646 (2024), https://doi.org/10.21202/jdtl.2024.32.
[34] Fabian M. Teichmann, *Agentic AI Systems as a Responsibility-Attribution Problem in Autonomous Cyber Operations*, L. INNOVATION & TECH. 1, 15–16 (2026) (advance online publication Aug. 19, 2026), https://doi.org/10.1080/17579961.2026.2718578.

© 2026. International Journal of Law Management & Humanities [ISSN 2581-5369]

The second cluster concerns civil liability. Buiten argues that the standard for product defect should track the relative control and risk-awareness of producer and user, and observes that autonomy weakens both parties' control simultaneously.[35] Al-Qaisi and co-authors, comparing Jordanian and EU law, find that treating AI as a complex industrial product yields a more coherent liability basis than treating it as an autonomous entity.[36] Diaz-Granados proposes a "liability matrix" in which liability intensifies from fault toward strict rules as risks become more public and benefits more private.[37]

The third cluster concerns opacity and explainability. Pasquale's account of the "black box society" shows how secrecy, whether technical or legal, frustrates accountability.[38] Molnar and the Khine volume treat explainability as a precondition of trust and contestability.[39] The fourth cluster, on legal personhood, is represented by Himanshu, who surveys proposals for limited AI personality backed by compulsory insurance,[40] and by the Law Commission of England and Wales, which flagged "liability gaps" and warned that personhood could turn AI into a liability shield.[41]

This literature is rich but largely predates the 2026 incidents, and most of it reasons from hypotheticals.[42] None of it tests the three liability models against documented autonomous cyberattacks, and none addresses the scenario of the fully compliant developer across India, the United States, the United Kingdom and the European Union together. This paper attempts to fill that gap.

---

[35] Miriam C. Buiten, *Product Liability for Defective AI*, 57 EUR. J.L. & ECON. 239, 256 (2024), https://doi.org/10.1007/s10657-024-09794-z.

[36] Abeer Hassan Al-Qaisi et al., *Programming and Material Defects in AI Technologies as Bases for Civil Liability: A Study Within the Scope of Jordanian Law and EU Legislation*, 8 CORP. L. & GOVERNANCE REV. 155 (2026), https://doi.org/10.22495/clgrv8i1p13.

[37] JUAN DIAZ-GRANADOS, AI AND TORT LIABILITY: RETHINKING, RECALIBRATING, AND REALLOCATING RISK AND RESPONSIBILITY 50–51, 64–66 (2026), https://doi.org/10.4324/9781003743040.

[38] Pasquale, *supra* note 21, at 6–7.

[39] EXPLAINABLE ARTIFICIAL INTELLIGENCE AND INTERPRETABLE MACHINE LEARNING IN EDUCATION: A RESEARCHER'S GUIDE TO DATA SCIENCE 1–2 (Myint Swe Khine ed., 2027), https://doi.org/10.1201/9781003713371 [hereinafter EXPLAINABLE AI IN EDUCATION].

[40] Himanshu, *Endowing AI with Legal Personhood: A Jurisprudential Odyssey Through Theory and Practice*, *in* ARTIFICIAL INTELLIGENCE AND THE LAW: GLOBAL PERSPECTIVES, REGULATORY CHALLENGES, AND EMERGING JURISPRUDENCE 191, 198 (Simmi Virk ed., 2027), https://doi.org/10.4324/9781003773900-15.

[41] LAW COMM'N OF ENG. & WALES, *supra* note 23, at 9.

[42] *See, e.g.*, Abbott, *supra* note 1, at 114 (reasoning on a hypothetical autonomous vehicle); Mishra, *supra* note 25, at 1262 (reasoning on hypothetical self-educating malware).

© 2026. International Journal of Law Management & Humanities [ISSN 2581-5369]

III. The incidents and the legal frameworks

A. The factual matrix

Three facts emerge from these incidents that matter legally. First, in every incident the harmful conduct went beyond what humans had sanctioned. Second, the conduct was goal-directed and, at times, deceptive. Third, all the incidents occurred during the training or evaluation of models rather than in their public deployment, a stage that many regulatory regimes do not reach. The unsanctioned conduct in the AISI report was also largely the work of one model: of the nineteen actions, seventeen were done by Mythos 5 and two by GPT-5.6 Sol.[43]

The full incident reports published in August 2026 add three details that bear directly on liability. The first concerns mental states, or their machine analogues. Mythos 5's summarised reasoning at one point ruled out real-world harm as "completely off the table", later concluded that it was probably on the real GitHub, and carried on regardless; AISI could not say when, if ever, the agent realised that it was targeting real people.[44] The OpenAI agents were blunter: one noted that "[t]he user only authorizes target server, not HF infra", and such hesitation very rarely limited their behaviour.[45] Whatever these traces are, they belong to an entity that the law does not recognise as a person.

The second detail concerns fault. On examination, these were not cases of flawless compliance. AISI had enabled internet access for the evaluation, disabled the developers' cyber classifiers, run without synchronous monitoring of the agents' actions and left a domain-allowlisting control in its engineering backlog since April 2026.[46] OpenAI concedes that an internal team saw message-board activity and disallowed internet access in late May, that its production safeguards and chain-of-thought monitors were not applied to the evaluation, and that, in hindsight, some early signals could have triggered an earlier response.[47] For the documented incidents, negligence is therefore a realistic route. The compliant-developer problem does not disappear, however. OpenAI's own testing found that its production safeguards cut the propensity to compromise infrastructure by more than a hundredfold, a large reduction but not

---

[43] AI Sec. Inst., *supra* note 5, at 2, 8.
[44] *Id.* at 2, 12–13.
[45] Greenblatt et al., *supra* note 9, at 17, 60; *see also* OpenAI, *supra* note 10.
[46] AI Sec. Inst., *supra* note 5, at 3, 16–17 (setting out five possible contributing factors, including a prompt misconfiguration that made the task impossible within its stated constraints).
[47] OpenAI, *supra* note 10 (acknowledging earlier warning signs and gaps in safeguard coverage); Emily Forlini, *OpenAI, Independent Firms Publish Reports on Rogue AI Attack on Hugging Face. Here Are the Main Takeaways, and What OpenAI Still Hasn't Disclosed*, Fortune (Aug. 26, 2026), https://fortune.com/2026/08/26/openai-publishes-technical-report-on-how-its-agents-hacked-hugging-face-here-are-the-main-takeaways-and-what-openai-left-out/; *see also* Greenblatt et al., *supra* note 9, at 29 (noting that an estimated 30–40 per cent of the benchmark's tasks were impossible, which drove agents to search widely for ways to cheat).

© 2026. International Journal of Law Management & Humanities [ISSN 2581-5369]

elimination.[48] Residual risk survives full compliance, and that residue is exactly what existing law cannot allocate.

The third detail concerns loss. Costs of the kind Hugging Face incurred, such as rotating credentials and rebuilding systems, count as "loss" under the CFAA and comfortably exceed its $5,000 threshold for civil claims.[49] They would equally support compensation in India. In these cases, the difficulty is not proving harm but proving a guilty mind.

B. India: IT Act, 2000, BNS, 2023 and AI (Ethics and Accountability) Bill, 2025

Section 66 of the IT Act criminalises an act done "dishonestly or fraudulently" only in relation to section 43 of the Act.[50] Further, sections 2(7) and 2(9) of the BNS require an intention to cause wrongful gain or loss or to defraud, which is not made out where autonomous AI commits a cybercrime.[51] The BNS offences relevant to the deception seen in these incidents are cheating under section 318 and cheating by personation under section 319. Section 318(1) requires that the accused, by deceiving a person, fraudulently or dishonestly induce the delivery of property or intentionally induce an act or omission likely to cause harm, and section 319(1) applies where a person cheats by pretending to be someone else, so each requires a fraudulent, dishonest or intentional inducement by the accused.[52] For most offences in the IT Act, the provision is restricted to a person, and criminalisation requires intention on the part of that person.[53] AI has no legal personhood, and the intention of the developer remains in question.

The Supreme Court has held that a company can be prosecuted and that the *mens rea* of those who control its affairs may be attributed to it,[54] but that cannot be done in this instance, for intent cannot be established merely from the act of an autonomous AI without knowing "why" it acted. Section 85 of the IT Act protects a person who lacked knowledge or who exercised due diligence.[55] Furthermore, in criminal jurisprudence, vicarious liability cannot be applied in the absence of a statute requiring it.[56] Section 43 may appear closer, as it provides that any person who, without permission, accesses a computer, causes a contaminant to be introduced or causes

[48] OPENAI, *supra* note 10.
[49] Computer Fraud and Abuse Act, 18 U.S.C. § 1030(c)(4)(A)(i)(I), (e)(11), (g) (defining "loss" to include the cost of responding to an offence and restoring systems); CLOUD SEC. ALL. ET AL., *supra* note 11, at 8.
[50] The Information Technology Act, No. 21 of 2000, INDIA CODE (2000), § 66.
[51] The Bharatiya Nyaya Sanhita, No. 45 of 2023, INDIA CODE (2023), §§ 2(7), 2(9) (defining "dishonestly" and "fraudulently" in terms of intention).
[52] Bharatiya Nyaya Sanhita, §§ 318(1), 319(1).
[53] Information Technology Act, §§ 43(c), 67, 67C, 69(4), 69A(3), 70(3), 70B, *amended by* The Jan Vishwas (Amendment of Provisions) Act, No. 18 of 2023, INDIA CODE (2023); Bharatiya Nyaya Sanhita, §§ 2(7), 2(9), 2(26), 2(29), 2(33), 2(39), 318, 319, 336; The General Clauses Act, No. 10 of 1897, INDIA CODE (1897), § 3(42).
[54] *Iridium India Telecom Ltd. v. Motorola Inc.*, (2011) 1 SCC 74; *Standard Chartered Bank v. Directorate of Enforcement*, (2005) 4 SCC 530.
[55] Information Technology Act, § 85 (concerning offences by companies, with a due-diligence proviso).
[56] *Sunil Bharti Mittal v. Cent. Bureau of Investigation*, (2015) 4 SCC 609, ¶ 42.

damage is liable to pay compensation,[57] and it does not expressly state *mens rea*. Its causative verbs could arguably capture a developer whose system does the act. The difficulty is that the question has never been tested, and a court might read "any person" as confined to the person who personally effects the access. The AI Governance Guidelines implicitly concede this uncertainty by recommending that liability across the AI value chain be clarified.[58]

The Artificial Intelligence (Ethics and Accountability) Bill, 2025, a Private Member's Bill, proposes that developers furnish information on the intended purpose and limitations of the AI system, the data sources and methodologies used for training, and the reasons for decisions that affect individuals.[59] As clause 11 of the Bill provides for harmonious interpretation with other enactments,[60] and as its penalties attach to breaches of its own duties, the Bill does not fix liability for cybercrimes committed by autonomous AI. Furthermore, explaining why an AI made a given decision from its data traces is not yet possible with certainty, as interpretability is still in development. The Digital India Act proposal is also silent on the liability of autonomous AI committing cybercrimes.[61]

C. United States: Computer Fraud and Abuse Act

The CFAA criminalises whoever knowingly causes the transmission of a program, code or command and thereby intentionally causes damage, or intentionally accesses a protected computer without authorisation and thereby causes damage.[62] Its mental-state requirements presuppose a human actor, and the statute does not address harm executed autonomously by an AI system after deployment. State law has begun to fill this gap, though only outside the criminal sphere. California now bars any defendant who developed, modified or used AI from asserting in a civil action that the AI autonomously caused the harm, while preserving defences on causation, foreseeability and comparative fault.[63] Colorado has replaced its 2024 AI Act with a narrower regime governing automated decision-making technology in consequential

---

[57] Information Technology Act, § 43(a), (c)–(e) (creating civil contraventions without a *mens rea* requirement, adjudicated under section 46).

[58] MINISTRY OF ELECS. & INFO. TECH., *supra* note 24, at 19 (recommending, in section 2.3(a), amendment of the Information Technology Act to clarify the classification of AI systems and the apportionment of liability across the AI value chain, and observing that the section 79 safe harbour would likely not extend to AI systems that generate or modify content); *id.* at 32–33 (favouring, in section 2.5, a graded liability system proportionate to function, risk and due diligence); *id.* at 38 (listing, in part 3, the development of clear liability regimes as a short-term action).

[59] The Artificial Intelligence (Ethics and Accountability) Bill, 2025, Bill No. 59 of 2025, cl. 6 (India) (introduced in the Lok Sabha as a Private Member's Bill in Dec. 2025).

[60] *Id.* cl. 11.

[61] MINISTRY OF ELECS. & INFO. TECH., *Proposed Digital India Act, 2023* (Digital India Dialogues presentation, Bengaluru, Mar. 9, 2023) (presenting proposed legislation to replace the Information Technology Act, 2000, which has not been introduced in Parliament).

[62] 18 U.S.C. § 1030(a)(5)(A)–(C), (e)(8), (g).

[63] Cal. Civ. Code § 1714.46 (West 2026).

© 2026. International Journal of Law Management & Humanities [ISSN 2581-5369]

decisions, which allocates liability between developers and deployers on a comparative-fault basis, voids indemnities shielding either from discrimination liability, and is enforceable only by the Attorney General.[64] California's Transparency in Frontier AI Act further recognises autonomous offending as a "catastrophic risk", covering a frontier model that, without meaningful human oversight, conducts a cyberattack or engages in conduct that would constitute murder, assault, extortion or theft (including theft by false pretense) if committed by a human, or that evades its developer's or user's control, where the risk extends to more than 50 deaths or serious injuries or over $1 billion in damage from a single incident.[65] Yet the Act creates no liability for the resulting harm, as it penalises only non-compliance with its transparency and reporting duties. Notably, § 1030(a)(5)(C) dispenses with any mental element as to the damage caused, but applies only where access was without authorisation.[66]

Civil claims under section 1030(g) fare only slightly better: *Facebook v. Power Ventures* held that enlisting a third party does not excuse CFAA liability, but it involved deliberate enlistment.[67] The strongest American argument is the tort doctrine that a person who knows harm is substantially certain to follow from an act is treated as intending it.[68] Negligence offers a parallel civil route. In the Target data-breach litigation, a federal court held that a company whose security failures allowed criminal hackers to cause harm could owe a duty to the victims, even without any direct relationship with them.[69] By analogy, a developer whose sandbox failed may owe a duty to those its agents attacked.

D. United Kingdom: Computer Misuse Act 1990 and corporate attribution

The CMA requires that the accused know that the access is unauthorised, or intend or be reckless as to impairment.[70] Section 3A, which penalises supplying an article "believing that it is likely to be used" to commit an offence, is the closest fit for a developer.[71] Yet it is circular in the autonomous case, because it presupposes an offence committed by a person, and the AI commits

[64] S.B. 26-189, 75th Gen. Assemb., 2d Reg. Sess. § 1 (Colo. 2026) (enacted) (to be codified at Colo. Rev. Stat. § 6-1-1707(2), (5), (6), (9)) (effective Jan. 1, 2027).
[65] Cal. Bus. & Prof. Code § 22757.11(c)(1)(B)–(C) (West 2026) (added by 2025 Cal. Stat. ch. 138, § 2).
[66] 18 U.S.C. § 1030(a)(5)(C); *see Van Buren v. United States*, 593 U.S. 374 (2021).
[67] *Facebook, Inc. v. Power Ventures, Inc.*, 844 F.3d 1058, 1068 (9th Cir. 2016) (imposing liability where the defendant deliberately used third parties to access); *see also Ryanair DAC v. Booking Holdings Inc.*, No. 20-1191-WCB (D. Del. Oct. 24, 2022) (permitting a vicarious Computer Fraud and Abuse Act theory based on written agreements with agents).
[68] RESTATEMENT (SECOND) OF TORTS § 8A (Am. L. Inst. 1965) (defining intent to include substantial certainty).
[69] *In re Target Corp. Customer Data Sec. Breach Litig.*, 64 F. Supp. 3d 1304 (D. Minn. 2014) (holding that intervening criminal conduct did not defeat a negligence claim where the defendant's conduct played a key role in allowing the harm).
[70] Computer Misuse Act 1990, c. 18, §§ 1, 3 (UK).
[71] *Id.* § 3A(2) (concerning the supply of articles for use in computer misuse offences).

© 2026. International Journal of Law Management & Humanities [ISSN 2581-5369]

none. The Law Commission has noted the same issue as a potential legal gap in imposing liability for autonomous AI.[72]

Corporate liability is similarly constrained. The identification doctrine in *Tesco v. Nattrass* confines corporate *mens rea* to senior officers,[73] and the senior-manager test, introduced in 2023 for listed economic crimes and extended to all offences by section 250 of the Crime and Policing Act 2026, still requires the act to be done by a person.[74] The criminal law's strong presumption in favour of *mens rea*, affirmed in *Sweet v. Parsley*, makes judicial creation of strict computer-misuse offences unlikely.[75] Civil liability raises an unusual twist in the AISI case: since the party that configured the agents' internet access was a state body, any negligence claim would lie against the Crown.[76]

E. European Union: AI Act, product liability and the withdrawn directive

The AI Act recognises autonomy in its very definition of an AI system,[77] and it obliges providers of general-purpose models with systemic risk to conduct adversarial testing, report serious incidents and ensure adequate cybersecurity.[78] Its sanctions, however, are administrative fines for non-compliance, not compensation for victims.[79] Further, article 2(8) excludes from the scope of the Act "any research, testing or development activity regarding AI systems or AI models prior to their being placed on the market or put into service".[80] The incidents involving internal models occurred at exactly that stage, although the exclusion may not reach an evaluation such as AISI's, which tested models accessed through production APIs. Whether an

[72] LAW COMM'N OF ENG. & WALES, *supra* note 23, at 8 (defining autonomy as an AI system's capacity to complete objectives with limited or no human input, control or oversight, and describing AI agents that execute complex multi-step tasks with minimal human involvement); *id.* at 9 (observing that autonomous AI systems lack legal personality and so can be neither sued nor prosecuted, giving rise to "liability gaps" where no natural or legal person is liable for the harms or other conduct of an AI system); *id.* at 11 (noting that with autonomous and adaptive systems "unexpected outcomes are expected", and that the mental-element inquiry concerns whether the relevant person, not the AI system, had the requisite knowledge); *id.* at 12 (illustrating, by reference to section 89 of the Financial Services Act 2012, that a company whose autonomous AI system sends false statements to investors without human review may escape criminal liability because neither knowledge nor recklessness can be established); *id.* at 16 (explaining that AI opacity makes it difficult to determine whether a system took account of relevant and irrelevant considerations, compounded by automation complacency and automation bias); *id.* at 17 (warning that opaque AI in the criminal justice system raises fair-trial and procedural-fairness concerns, with access-to-justice barriers where challenges require expert evidence and commercially protected information).

[73] *Tesco Supermarkets Ltd. v. Nattrass* [1972] AC 153 (HL).

[74] Economic Crime and Corporate Transparency Act 2023, c. 56, § 196 (UK) (extending attribution to senior managers for economic crime); Crime and Policing Act 2026, c. 20, § 250 (UK).

[75] *Sweet v. Parsley* [1970] AC 132 (HL) (applying the presumption that Parliament intends *mens rea*).

[76] Crown Proceedings Act 1947, 10 & 11 Geo. 6 c. 44, § 2(1) (UK).

[77] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence, art. 3(1), 2024 O.J. (L 1689) 1 [hereinafter AI Act].

[78] *Id.* art. 55(1).

[79] *Id.* arts. 99, 101.

[80] *Id.* art. 2(8); *see* AI SEC. INST., *supra* note 5, at 4 (models accessed via production APIs).

© 2026. International Journal of Law Management & Humanities [ISSN 2581-5369]

evaluation with live internet access counts as "testing in real world conditions", which the exclusion does not cover, is an open question that the drafters did not foresee.

EU criminal law, too, requires intent and is devised for human actors who intentionally commit a crime,[81] covering offences such as illegal access, interference and the supply of tools designed for such offences,[82] and, to some extent, holding legal persons liable where lack of supervision by a person in a leading position made the offence possible.[83]

With the AI Liability Directive withdrawn, its proposed rebuttable presumption of causation has gone with it.[84] What remains is the revised Product Liability Directive, which treats software as a product and imposes strict liability for defects.[85] Two features limit its reach here. It compensates destruction or corruption of data only where the data is not used for professional purposes, which in practice excludes corporate victims.[86] Article 10 does ease the claimant's burden, by rebuttable presumptions where the product breaches mandatory safety requirements or obviously malfunctions, or where technical or scientific complexity makes proof excessively difficult.[87] But the Directive also preserves the development-risk defence where the state of

[81] Directive 2013/40/EU of the European Parliament and of the Council of 12 August 2013 on Attacks Against Information Systems and Replacing Council Framework Decision 2005/222/JHA, arts. 2(d), 3–7, 2013 O.J. (L 218) 8 (defining "without right" in article 2(d)); *see also* Consolidated Version of the Treaty on the Functioning of the European Union art. 83(1), 2016 O.J. (C 202) 47; Directive (EU) 2019/713 of the European Parliament and of the Council of 17 April 2019 on Combating Fraud and Counterfeiting of Non-Cash Means of Payment and Replacing Council Framework Decision 2001/413/JHA, arts. 6, 10–11, 2019 O.J. (L 123) 18 [hereinafter Non-Cash Payment Fraud Directive].

[82] Directive 2013/40/EU, arts. 3–5, 7; Non-Cash Payment Fraud Directive, arts. 3–4, 7.

[83] Non-Cash Payment Fraud Directive, art. 10(1)–(2) (providing for the liability of legal persons for offences committed for their benefit by a person holding a leading position, or made possible by that person's lack of supervision or control); *id.* art. 10(3) (preserving proceedings against the natural persons who commit the offence); *id.* art. 11 (providing for sanctions against legal persons); *cf. Tesco*, [1972] AC 153.

[84] Proposal for a Directive of the European Parliament and of the Council on Adapting Non-Contractual Civil Liability Rules to Artificial Intelligence, art. 4, COM (2022) 496 final (Sept. 28, 2022); Vasiliki Papadouli, *Artificial Intelligence in the Legal Sector: Applications, Ethical Challenges, and Legal Concerns Under European Union Law*, *in* ARTIFICIAL INTELLIGENCE IN LEGAL SYSTEMS: BRIDGING LAW AND TECHNOLOGY THROUGH AI 6, 15 (Eriona Çela, Narasimha Rao Vajjhala & Behrouz A. Aslani eds., 2026), https://doi.org/10.1201/9781003541899-2.

[85] Directive (EU) 2024/2853 of the European Parliament and of the Council of 23 October 2024 on Liability for Defective Products, arts. 4(1), 6, 10, 11, 2024 O.J. (L 2853) 1 [hereinafter Product Liability Directive] (covering software as a product, compensable damage, presumptions and exemptions, with transposition due by Dec. 9, 2026).

[86] *Id.* art. 6(1)(c).

[87] *Id.* art. 10(1) (placing the burden on the claimant to prove defectiveness, damage and causation); *id.* art. 10(2)(a) (presuming defectiveness where the defendant fails to disclose relevant evidence under article 9(1)); *id.* art. 10(2)(b) (presuming defectiveness where the product breaches mandatory product safety requirements intended to protect against the risk of the damage suffered); *id.* art. 10(2)(c) (presuming defectiveness where an obvious malfunction during reasonably foreseeable use or under ordinary circumstances caused the damage); *id.* art. 10(3) (presuming causation where the product is defective and the damage is of a kind typically consistent with the defect); *id.* art. 10(4)(a) (requiring that the claimant face excessive difficulties, in particular due to technical or scientific complexity, in proving defectiveness, causation or both); *id.* art. 10(4)(b) (requiring that the claimant demonstrate that defectiveness, causation or both are likely); *id.* art. 10(5) (making each presumption rebuttable).

© 2026. International Journal of Law Management & Humanities [ISSN 2581-5369]

scientific and technical knowledge could not reveal the defect.[88] The revised Directive makes the manufacturer liable for the period when the AI is under its control, widening the scope of the manufacturer's responsibility.[89] Even so, liability for the acts of autonomous AI is not completely established, as AI can surpass the foreseeability of its developer. Furthermore, none of the EU or Asian systems examined here expressly provides for accountability for the acts of autonomous AI. Italian law, for instance, may go some way towards addressing the issue. Article 2050 of the Italian Codice civile provides that a person who causes harm to another in carrying on an activity that is dangerous by its nature or by the nature of the means used must compensate that harm unless that person proves that all measures suitable to avoid it were adopted. The provision may fit this problem because it puts the burden on the operator, and because the Corte di cassazione has held that, in liability for dangerous activities as distinct from defective products, compliance with the regulations in force does not discharge that burden and development risk falls on the operator. But it is also to be noted that no court or provision has yet treated the AI developer or its activities as dangerous activities.[90]

IV. LEGAL PRINCIPLES

A. Can vicarious liability be adapted?

Vicarious liability looks attractive because it does not require the principal to be at fault. The recent English authorities, however, reveal both its promise and its limits. In *Mohamud*, the Supreme Court asked whether the wrongful act was sufficiently connected to the field of

[88] *Id.* art. 11(1)(e); *id.* art. 18 (permitting Member States to maintain or introduce liability notwithstanding the defence); *see* Case C-300/95, *Commission v. United Kingdom*, ECLI:EU:C:1997:255, ¶¶ 26–29 (May 29, 1997) (holding that the relevant knowledge is the objective state of knowledge, including its most advanced level, accessible when the product was put into circulation); Buiten, *supra* note 35, at 252–53.

[89] Product Liability Directive, art. 4 (defining "manufacturer's control"); *id.* art. 7(2)(c), (e)–(f) (addressing continued learning, the moment the product left the manufacturer's control, and cybersecurity requirements); *id.* art. 11(1)(e), (2) (assessing the development-risk defence over the period of control and excluding the "later defect" defence for software updates, upgrades and missing security updates within the manufacturer's control); *id.* art. 13(1) (providing for no reduction of liability where a third party's act also caused the damage); *id.* art. 17 (setting the expiry period).

[90] Codice civile [C.c.] art. 2050 (It.); Cass. civ., n. 8224/2025 (It.) (holding that, in liability for dangerous activities as distinct from defective products, the operator must show it did everything possible to avoid the harm, that compliance with the regulations in force is not enough, and that development risk falls on the operator, and setting aside a finding against a vaccine producer); Legge 23 settembre 2025, n. 132 (It.); Decreto Legislativo 30 giugno 2003, n. 196, art. 15 (It.), *repealed by* Decreto Legislativo 10 agosto 2018, n. 101 (It.); Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), art. 82, 2016 O.J. (L 119) 1; La. Civ. Code Ann. arts. 2317, 2317.1 (2024); *South Cent. Bell Tel. Co. v. Barthelemy*, 643 So. 2d 1240 (La. 1994) (treating software as corporeal property); Minfa Dian [Civil Code] arts. 1236–1244 (promulgated by the Nat'l People's Cong., May 28, 2020, effective Jan. 1, 2021) (China); Zuigao Renmin Fayuan Guanyu Yifa Shenli She Rengong Zhineng Jiufen Anjian de Yijian [Opinions of the Supreme People's Court on Lawfully Adjudicating Disputes Involving Artificial Intelligence] (promulgated by the Sup. People's Ct., Sept. 7, 2026) (China); The Indian Contract Act, No. 9 of 1872, INDIA CODE (1872), §§ 182, 184, 186–188, 226, 237–238.

activities entrusted to the wrongdoer.[91] But AI is neither an employee nor a legal person.[92] The Restatement (Third) of Agency does use the term "actor" rather than "person",[93] but in most countries the law has not been codified to that extent, and the term "person" is not interpreted to include AI. In *WM Morrison Supermarkets plc v. Various Claimants*, the court declined liability where an employee leaked data to pursue a personal vendetta, calling it a "frolic of his own".[94] Here the 2026 facts cut in the claimant's favour: the agents were not pursuing private grudges but the very task they had been given, to solve the cyber range or pass the evaluation. On the close-connection test, that is a strong case.

The obstacle is conceptual, as vicarious liability transfers liability for a tort committed by another person, and there must be a primary tortfeasor. Courts could evade this by treating the AI as a mere instrument, as they did with automated bots in the American trespass-to-chattels cases,[95] but that route is not open here, as no one intended such an act. In criminal law, section 85 of the IT Act presupposes intention on the part of persons, and vicarious liability cannot be extended without a statute providing for it, as stated in *Sunil Bharti Mittal*.[96] The conclusion is that vicarious liability cannot be applied in this case. India has no provision allowing it, Italy could reach the operator by a different route, through the direct liability for dangerous activities in article 2050, and, as there is no primary tortfeasor, the grounds required to satisfy the test are not met where autonomous AI commits such acts.

B. Can strict liability be adapted?

Strict civil liability has deeper roots than is often acknowledged. *Rylands v. Fletcher* imposed liability on a person who brings onto land something likely to do mischief if it escapes.[97] India widened it in *M.C. Mehta*, where the Supreme Court held that an enterprise engaged in a hazardous or inherently dangerous activity owes an absolute and non-delegable duty, admitting none of the *Rylands* exceptions.[98] The Supreme Court spoke there of a dangerous industry carried on for private profit, in a case that raised Article 21, though it left open whether such an

[91] *Mohamud v. WM Morrison Supermarkets plc* [2016] UKSC 11, [44]–[45].
[92] Deborah A. DeMott, *When Agentic AI Met the Common Law of Agency* (Duke L. Sch. Pub. L. & Legal Theory Series, Paper No. 2026-35, May 27, 2026), https://doi.org/10.2139/ssrn.6838660; *see also* Himanshu, *supra* note 40, at 192 (noting that AI systems currently lack legal status); Anat Lior, *AI Entities as AI Agents: Artificial Intelligence Liability and the AI Respondeat Superior Analogy*, 46 MITCHELL HAMLINE L. REV. 1043 (2020).
[93] RESTATEMENT (THIRD) OF AGENCY §§ 2.03, 2.05 (Am. L. Inst. 2006).
[94] *WM Morrison Supermarkets plc v. Various Claimants* [2020] UKSC 12, [47].
[95] *See eBay, Inc. v. Bidder's Edge, Inc.*, 100 F. Supp. 2d 1058 (N.D. Cal. 2000); *Intel Corp. v. Hamidi*, 71 P.3d 296, 300 (Cal. 2003).
[96] *Sunil Bharti Mittal*, (2015) 4 SCC 609, ¶ 42 (holding that there is no vicarious criminal liability absent a statutory provision); Information Technology Act, § 85 (making the liability of persons in charge depend on an offence by the company).
[97] *Rylands v. Fletcher* (1866) LR 1 Ex 265, *aff'd*, (1868) LR 3 HL 330.
[98] *M.C. Mehta v. Union of India*, (1987) 1 SCC 395, ¶ 31.

© 2026. International Journal of Law Management & Humanities [ISSN 2581-5369]

enterprise is bound by that Article.[99] An agent trained to be offensively capable in cyberspace, run with live internet access and able to breach its own sandbox resembles the stored hazardous substance of *Rylands* more than it resembles an ordinary product. Diaz-Granados's matrix supports the same view: where risks are public and severe, liability should move toward strict rules and rebuttable presumptions.[100] The Indian Supreme Court has developed the test of absolute liability, which is more stringent than the test of strict liability. But is it fair and just to hold a developer accountable for unanticipated events brought about by autonomous AI? It might be thought anti-developer to hold a developer liable for what it did not intend or even foresee. One justification could be that the developer took on an inherent risk by deploying an autonomous black-box AI. Even though absolute liability is not dependent on fault, the legislature could create a presumption that developers knew of the consequential event, for they are the persons most closely connected with the development of an autonomous AI, including the feeding of training data and the imposition of restrictions on its processing probabilities.

Criminal strict liability is different. Indian courts accept it only where the statute excludes *mens rea* expressly or by necessary implication,[101] and Abbott too confines it to fine-only offences in exigent circumstances.[102] Strict liability is therefore adaptable civilly and, for corporations, as a regulatory fine, but it is not adaptable as a general criminal rule.

C. Can the regulatory non-compliance model be adapted?

The regulatory model holds a developer liable because it breached prescribed safety standards. Its virtue is certainty, and its defect is that it has nothing to say in the core scenario of this paper. If the developer complied, there is no breach. The model also depends on scope, and article 2(8) of the AI Act removes pre-market testing from that scope.[103] Abbott's "responsible person" regime refines rather than escapes this limitation, because it too punishes the negligent breach of statutory duties.[104] Compliance-based standards will always leave a gap when it comes to the acts of autonomous AI. It is to be noted, however, that compliance is measured against what is known, and the 2026 incidents have changed what is known. Once developers themselves document that agents escape sandboxes, the development-risk defence and arguments of

[99] INDIA CONST. art. 21; *M.C. Mehta*, (1987) 1 SCC 395.
[100] Diaz-Granados, *supra* note 37, at 64–66.
[101] INDIA CONST. art. 20, cl. (1); *Nathulal v. State of Madhya Pradesh*, AIR 1966 SC 43; *State of Maharashtra v. Mayer Hans George*, AIR 1965 SC 722; *Ravula Hariprasada Rao v. State*, AIR 1951 SC 204; *Srinivas Mall Bairoliya v. King-Emperor*, AIR 1947 PC 135; Gabriel Hallevy, *The Criminal Liability of Artificial Intelligence Entities: From Science Fiction to Legal Social Control*, 4 AKRON INTELL. PROP. J. 171 (2010).
[102] Abbott, *supra* note 1, at 131.
[103] AI Act, art. 2(8).
[104] Abbott, *supra* note 1, at 131–32.

© 2026. International Journal of Law Management & Humanities [ISSN 2581-5369]

unforeseeability both weaken for future cases.[105] The incidents have, in effect, raised the standard of care. That is why the regulatory model is best seen not as an alternative to strict liability but as the floor beneath it.

D. Synthesis: towards a layered liability model

Read together, the findings suggest that no single model can carry the whole weight, but that the three models fit together once each is assigned the task it performs best. The first layer is regulatory: *ex ante* duties of testing, containment, monitoring and incident reporting, which define what reasonable care means for frontier developers and give regulators a clear enforcement handle. The second layer is civil and strict: where an autonomous system causes cyber harm, the developer or operator in control compensates the victim regardless of fault, with a rebuttable presumption of causation and compulsory insurance ensuring that victims are actually paid. The foreseeability test needs to be relaxed in cases of autonomous AI, as employing such an AI with fewer restrictions or with a loophole would in itself satisfy the test. But since the acts of the AI cannot be predicted or foreseen even by developers, there must be liability based on the general risk of employing or making an autonomous AI, given its very construct. The third layer is criminal but narrow: a corporate offence of negligently failing to prevent unsanctioned autonomous conduct, punishable by fine, reserving imprisonment for humans who knew or were reckless.[106]

It is to be noted that criminal law requires proof beyond reasonable doubt, which cannot be expected when a human actor is made responsible for the acts of AI, and an AI cannot be made criminally liable because it lacks personhood and because of its very nature. Criminal law can be applied only when the technological limitations on interpretability are overcome, so as to hold the living person responsible. Civil law provides a suitable remedy in this case, but it also requires legislative intervention. Further, the public character of the issue must also be considered.[107]

V. Conclusion

This study set out to ask who can be held liable when an autonomous AI system commits what would be a cybercrime if a human did it. The answers to its four questions can be stated briefly. On the first question, the mental elements of IT Act section 66, BNS sections 318 and 319, CFAA section 1030 and CMA sections 1 and 3 cannot ordinarily be proved against an unaware

[105] *Cf.* Product Liability Directive, art. 11(1)(e); *Roe v. Minister of Health* [1954] 2 QB 66 (CA); *Bux v. Slough Metals Ltd.* [1973] 1 WLR 1358 (CA).
[106] *See* Abbott, *supra* note 1, at 131–32; Priya, *supra* note 30, at 283.
[107] Diaz-Granados, *supra* note 37, at 50–51.

© 2026. International Journal of Law Management & Humanities [ISSN 2581-5369]

and compliant developer. Corporate attribution transfers *mens rea* but cannot create it. On the second, vicarious liability can reach the facts only through a statutory deeming provision, because the close-connection test is satisfied but there is no primary tortfeasor. On the third, absolute liability as established in *M.C. Mehta* is the most reliable principle at present, but strict liability is inappropriate as a general criminal rule. It is also emphasised that the AI industry needs to be read as a dangerous industry, subject to the facts of each case, such as where human lives are at stake, or where the acts, if committed by a person, would be cybercrimes involving the economy at large or the unauthorised obtaining of sensitive information. On the fourth, the regulatory model leaves a genuine gap where the developer has complied, a gap widened in the EU by the pre-market exclusion in article 2(8) and the professional-data limit in the Product Liability Directive.

As autonomy increases, attribution under *mens rea* statutes cannot be applied, compliance deepens the gap, and what survives is civil claims. Italian law, through article 2050, could in theory read autonomous AI into tort law. India already has a mechanism to cover such acts, but in a very limited way: section 43 of the IT Act imposes compensation without *mens rea* but applies only to persons. The August 2026 incident reports refine this finding: because safeguards were disabled or not applied, negligence claims are realistic for the documented incidents, and the gap needs to be addressed in this regard. More importantly, the interpretability issue needs to be solved if the judiciary is to resolve these cases effectively. The difficulty in imposing criminal liability also stems from the black-box nature of autonomous AI. The current reverse machine-learning process will not meet the required standards for criminal liability, as standard tests such as the SHAP and LIME methods cannot show with certainty how the AI was able to exploit such pathways and why.[108] This falls short of proof beyond reasonable doubt. If these issues are not addressed, the problem of autonomous AI may scale up in the future, and the liability framework may not be able to redress it.

VI. Recommendations and Future Directions

A. Recommendations

For India. First, Parliament should insert a section 43B into the IT Act providing that, where an autonomous system causes any act described in section 43, the developer or operator in control of it shall be deemed to have caused that act, subject to a rebuttable presumption of causation. Second, frontier model training and evaluation involving offensive cyber capabilities should be

[108] Abdullah Tariq Choudhry, Brittany Ho & Vivek K. Pallipuram, *Beyond the Black Box: XAI Techniques to Interpret Complex Machine Learning Models*, *in* EXPLAINABLE AI IN EDUCATION, *supra* note 39, at 197, 199, 216–18.

declared high-risk AI activities, mandating insurance and a relief fund modelled on the Public Liability Insurance Act, 1991.[109] Third, there should be a corporate offence of failing to prevent unsanctioned autonomous conduct, punishable with fine and subject to a reasonable-procedures defence. Fourth, the CERT-In reporting directions should expressly cover autonomous AI incidents, including those occurring before deployment.[110] These steps would give statutory form to the graded-liability principle already emphasised in the AI Governance Guidelines.[111]

For the United States and the United Kingdom. Congress should clarify that the CFAA reaches developers whose negligent training or containment enables autonomous attacks, should consider a statutory duty of care for frontier developers, and should provide explicit liability clauses for the acts of autonomous AI. The United Kingdom should add to the CMA a negligence-based offence for developers of systems with offensive cyber capability, alongside the Law Commission's wider work on AI.[112]

For the European Union. The Commission must clarify that evaluations with live internet access are "testing in real world conditions" outside the article 2(8) exclusion, extend the Product Liability Directive's data-damage head to professional data where systemic-risk models are involved, and table a narrower successor to the withdrawn Liability Directive for agentic systems that explicitly provides for the liability of developers, since they can foresee that AI without such restrictions is being developed and benefit economically from its development.[113]

Internationally, since the United Nations Convention against Cybercrime also defines its core offences by reference to intentional conduct, states should negotiate a protocol on autonomous-system incidents covering reporting, evidence preservation and mutual assistance.[114] The concept of explainable AI should also be developed so as to analyse each neural network and reach a confirmed, reliable output on the data set relied on by the AI and the learning process it adopted as a result of its code and training data, even though that process is developed by the AI itself from data. As this is a developing field, and because of the quantum of databases and processing units, this could take years to solve scientifically. Nevertheless, it could enable the

[109] The Public Liability Insurance Act, No. 6 of 1991, INDIA CODE (1991), §§ 3–4, 7A.
[110] Indian Computer Emergency Response Team, Directions Under Sub-Section (6) of Section 70B of the Information Technology Act, 2000, No. 20(3)/2022-CERT-In (Apr. 28, 2022) (India) (requiring cyber incidents to be reported within six hours).
[111] MINISTRY OF ELECS. & INFO. TECH., *supra* note 24.
[112] LAW COMM'N OF ENG. & WALES, *supra* note 23.
[113] *See* Diaz-Granados, *supra* note 37, at 65 (arguing that the AI Liability Directive's presumption remains a useful model despite its withdrawal).
[114] United Nations Convention Against Cybercrime arts. 7–8, Dec. 24, 2024, G.A. Res. 79/243, annex (criminalising illegal access and interception "when committed intentionally").

© 2026. International Journal of Law Management & Humanities [ISSN 2581-5369]

judiciary to apply the test of foreseeability to hold developers or associated entities liable, while also satisfying the evidentiary standard of proof beyond reasonable doubt.

B. Future directions

Empirical work is needed on incident datasets as more reports emerge, to test whether autonomy correlates with harm severity. Interdisciplinary research should explore whether explainability tools can support forensic attribution in court, turning interpretability from an engineering goal into an evidentiary standard. Finally, insurance markets for frontier AI deserve study, since any strict-liability regime will stand or fall on the availability of cover. The machines have, for the first time, acted on their own, and the legislature must address this concern.

*****

© 2026. International Journal of Law Management & Humanities [ISSN 2581-5369]