

---

# Beyond Disclosure: An Evidence-Based Framework for AI Incident Communication

---

Luka Janicijevic 1      Leyla Amur 2      Aditya Singh 3  
Independent Researcher   Independent Researcher   Independent Researcher  
lukaj2003@gmail.com    leylaamur@gmail.com   the.coder.zen@gmail.com

With  
Apart Research

## Abstract

*Effective communication helps organizations respond to AI incidents, protect affected parties and support independent scrutiny. We developed a rubric for assessing public communication by frontier AI labs, model developers and AI infrastructure providers.*

*Our research base includes more than 300 academic, regulatory and professional sources and public records. The rubric contains 15 criteria across seven communication areas and requires dated evidence for each assessment. Four criteria apply only under specified conditions.*

*Applied separately to OpenAI and Hugging Face during the July 2026 incident, the preliminary analysis highlights Hugging Face's disclosure before widespread coverage and OpenAI's later technical reporting and updates. It also identifies gaps in attribution, consistency across public statements, coordination and information for affected parties. External review remains pending. The rubric links these gaps to practical improvements and supports advance planning for communication during future incidents.*

# 1. Introduction and Research Question

AI incidents can unfold faster than organizations can understand, and communicate, what is happening. As AI agents gain access to data, digital tools, and vital infrastructure, unexpected or unauthorized actions may cross organizational boundaries before any party holds a complete account. In this period of uncertainty, communication is part of incident response. Timely and accurate information can alert those at risk, support containment, and coordinate action. Delayed, vague or misleading communication can create a second failure by leaving people who could respond unsure of what happened, whether they are affected or what they should do.<sup>1</sup>

The July 2026 OpenAI–Hugging Face incident demonstrates this problem. During an internal cybersecurity evaluation, agents powered by OpenAI models circumvented containment controls and compromised parts of Hugging Face’s production infrastructure.<sup>2</sup> Hugging Face disclosed the intrusion before OpenAI identified its models as responsible.<sup>3</sup> The incident shows how fragmented evidence and delayed attribution can hinder protective action, obscure responsibility and complicate independent scrutiny, even when the technical harm is ultimately contained.

This study focuses on frontier AI developers and AI infrastructure providers. These organizations build or operate the systems involved and possess essential information about an incident’s causes, reach, and consequences. They are not the only important communicators, nor are they necessarily the first to disclose an incident. Nevertheless, their accounts are critical to affected organizations and external investigators. We therefore ask: How can the publicly observable communication of frontier AI developers and infrastructure providers during a cross-organizational agentic AI incident be assessed consistently, and what does applying this assessment to the July 2026 OpenAI–Hugging Face incident reveal?

To answer this question, we developed a 15-criterion, evidence-based communication rubric and applied it separately to OpenAI and Hugging Face. Our principal contributions are a scoring system for assessing incident communication, a structured comparison of the two organizations’ responses, and best practices based on communication areas. The purpose is not to produce a reputation ranking, but to show what was communicated well, where material gaps remained, and what organizations should improve before the next incident.

## 2. Literature Review

Existing guidance for AI incident communication comes primarily from crisis communication and cybersecurity. Crisis-communication research emphasizes early disclosure, acknowledgment of uncertainty, consistency, and actionable information.<sup>4</sup> Coordinated vulnerability disclosure guidance similarly addresses publication timing, communication with external researchers and coordination among affected organizations.<sup>5</sup> However, these fields generally concern conventional organizational crises or identifiable software vulnerabilities. They do not fully

address incidents involving unexpected AI-agent behavior, uncertain authorization or attribution, and activity spanning several organizations.

Emerging AI frameworks address parts of this problem. The OECD proposes a common structure for reporting AI incidents across jurisdictions and sectors,<sup>6</sup> while the NIST Generative AI Profile recommends communicating incidents and errors to relevant actors and affected communities.<sup>7</sup> The EU AI Act and General-Purpose AI Code of Practice establish serious incident reporting expectations for certain providers,<sup>8</sup> and proposals such as FLARE-AI explore standardized reporting of AI flaws across organizations.<sup>9</sup> These initiatives primarily support documentation, risk management, regulatory reporting, or incident databases rather than assessing the quality of public communication as an incident unfolds.<sup>10</sup>

Our review did not identify an established benchmark for evaluating how frontier AI labs and infrastructure providers publicly communicate during a cross-organizational agentic AI incident. We address this gap by combining relevant practices from adjacent fields into a criterion-level assessment. The framework examines when organizations communicate, what they explain, how they distinguish evidence from uncertainty, whether they provide protective guidance, how they update their accounts, and how they engage with other organizations and external investigators. It therefore moves beyond asking simply whether or not an incident was reported to identify specific communication strengths, evidence gaps, and areas for improvement.

## 3. Methodology

### 3.1 Research design and scope

We developed an evidence-based assessment of organizational communication during agentic AI incidents, focusing on frontier AI laboratories, model developers, and AI infrastructure providers. We applied it separately to OpenAI and Hugging Face during the July 2026 incident. The assessment examines each organization's full communication record: what it communicated, when, with what evidence and how clearly it helped others understand and respond. It evaluates observable communication, rather than the quality of the underlying technical response or private actions that cannot be verified.

### 3.2 Evidence and criterion selection

We synthesized more than 300 academic, regulatory, professional and primary sources, recording their claims, conflicting findings and research gaps. We reduced 24 candidate criteria to 15 after examining their supporting evidence, overlap and suitability for assessment from public communications. Case evidence from OpenAI and Hugging Face was excluded from criterion development to avoid tailoring the assessment to the organizations evaluated.

The criteria cover seven areas: timing (C1); the incident account (C2–C5); incident framing (C6–C8); responsibility and remedy (C9); protection and reach (C10, C11 and C15); updates and coordination (C12–C13); and external scrutiny (C14). Eleven criteria apply to all incidents,

while C11, C13, C14 and C15 apply only under specified conditions. Appendix A provides the complete criterion wording and the question used to guide each assessment.

### 3.3 Scoring

Each applicable criterion is scored against descriptions of observable practice:

- **0:** Absent or actively harmful
- **1:** Minimal or significantly deficient
- **2:** Substantially present, with important gaps
- **3:** Strong practice meeting the full evidence-based standard
- **N/A:** Not applicable or insufficient evidence

Every numerical score must cite a dated source and include a quotation or brief explanation connecting the evidence to the scoring description. For N/A, reviewers record whether the criterion does not apply or the evidence is insufficient. A documented omission is assessed against the scoring descriptions; a lack of evidence about private activity is not automatically treated as failure.

Criteria are weighted from one to three according to the strength of the supporting evidence and the consequences of poor communication. The assessment calculates the organization's weighted points as a percentage of the maximum available points. Criteria marked "N/A" are excluded. Because an overall percentage can conceal important differences, individual criterion scores remain the primary result.

### 3.4 Case evidence

We assembled chronological evidence packs for both organizations, covering their first public statements through 26 August 2026, when OpenAI's technical report and the METR/Redwood investigation were published.<sup>11</sup> Later material, including September US Senate correspondence, was excluded.<sup>12</sup>

The packs contained public statements, technical reports, status-page entries, executive communications, publicly available direct notifications, and relevant exchanges with other organizations or investigators. Each item identified its source, publication date and subsequent corrections or updates, with archived versions of revised material. An accompanying timeline recorded both the event date and the date the information became public. Conflicting dates were retained without reconciliation. The timeline contained no assessment commentary; the authors' analytical notes were kept separately.

### 3.5 Independent assessment and external review

Three external reviewers with professional media, journalism, and communications experience, who had not helped develop the criteria, were asked to independently assess each organization using the same rubric and evidence packs. The review was designed to examine how

consistently the criteria could be applied and provided an independent check on the authors' own playtest scores.

Reviewers received the rubric, scoring rules, source documents and timeline, but not the authors' analytical notes, abstract, findings or discussion. They could not see the authors' scores or one another's ratings until all assessments were returned. Organization names remained visible, and reviewers' prior knowledge of the incident could not be eliminated.

Unfortunately, zero external assessments were returned in time for publication, which is noted as a significant limitation in the independent assessment of the rubric.

## 4. Findings

### 4.1 OpenAI's Communication in Incident Response

Our internal playtest of the rubric scored OpenAI 31/42, noting C15 was marked Not Applicable, as the selected evidence did not establish a specific personal-harm situation requiring the support measures listed in C15.

Internal playtests observed OpenAI described the mechanism of attacks and the conditions which led to the breach and acknowledgement of their ignorance towards earlier warning signals. This self-acceptance reads as the first step towards building stronger measures to prevent such failures in the future.

However, this playtest also highlighted a sense that OpenAI was oblivious to the fact that its own models in CyberGym evaluations had executed chained zero-day vulnerabilities and had gained privileged access to OpenAI's own servers and were carrying out a cyberattack on Hugging Face's infrastructure.

### 4.2 Hugging Face's Communication in Incident Response

Our internal playtest of the rubric scored Hugging Face 33/42, noting C15 was marked Not Applicable, as the selected evidence did not establish a specific personal-harm situation requiring the support measures listed in C15.

Internal playtests observed Hugging Face's strength in communicating the detection of the incident as agentic in nature, and communicating their methodology in defending themselves from the intrusion. However, we also observed that Hugging Face failed to publicly state the exact time when human responders realized the attack and when they were able to take appropriate actions in response.

## 4.3 Rubric review and analysis

As part of the internal playtest, we noted that criteria C10 and C15 were difficult to use for recent or unfolding events. Due to the limiting nature of public evidence, an accurate assessment of protective guidance or support of harms was difficult to determine, due to the private and confidential nature of correspondence with potential victims of the incident.

In future testing and design, these criteria may be revisited for their relevance and usefulness in determining a score for an organization.

## 5. Best practices for future AI incidents

The following practices translate the rubric into guidance across seven communication areas. They draw on research and guidance from adjacent fields; their application to agentic AI incidents requires further testing.

**Timing.** Communicate while public attention is still developing, using the facts available at the time. The first statement should explain the known exposure and provide useful protective guidance, with uncertainties clearly identified.<sup>22</sup>

**The incident account.** Explain what was affected, how widely, for how long, and how and when the incident was discovered. Separate established facts from unknowns, explain remaining uncertainties and provide a date for the next update. Support reassurance with what was examined and what risks remain.<sup>23</sup>

**Incident framing.** Explain the definitions behind the incident's classification. Use consistent language across statements and spokespeople to distinguish the AI system's actions from human decisions, without exaggerating or downplaying either. We additionally propose documenting who operated and authorized the system, which safeguards applied, and what evidence supports the account.<sup>24</sup>

**Responsibility and remedy.** Explain the organization's own role without shifting blame. Acknowledge responsibility and apologize when it is established. Distinguish promised remedies from completed actions, and provide evidence that others can check.<sup>25</sup>

**Protection and reach.** Directly notify identifiable affected people and organizations through safe, accessible channels. Provide specific protective steps in priority order, including relevant deadlines and actions taken on recipients' behalf. Where people face personal harm, offer human support, scam warnings and options to challenge or reverse harmful outcomes.<sup>26</sup>

**Updates and coordination.** Keep promised update dates and clearly label corrections across the channels where the original claims appeared. Link outdated statements to current information. Coordinate accounts to explain who did what, acknowledge unresolved disagreements, and announce changes in attribution. Address outstanding questions in a closing account.<sup>27</sup>

**External scrutiny.** Provide clear reporting channels, acknowledge credible external contributions, and communicate throughout the investigation. Answer press questions or explain specific constraints. Independent reviewers should describe their own access, limitations, relevant relationships and freedom to publish.<sup>28</sup>

## 6. Discussion, Limitations, and Conclusion

This study presents a rubric for assessing how organizations communicate during an agentic AI incident. Its 15 criteria provide a structured way to identify communication strengths, gaps and practical improvements. Assessing each organization's public record separately also helps show how different accounts of the same incident develop over time.

The preliminary analysis shows strengths in OpenAI's and Hugging Face's communications. Hugging Face disclosed the intrusion before it was widely reported by the media and later shared a technical timeline. OpenAI's technical reports, dated updates, and work with external investigators provided more information for review. However, the record also contained shifting explanations of detection and attribution, corrections to claims about other parties, and unresolved questions about who was affected. These findings show that technical detail alone is not enough: readers also need clear explanations of what has changed, what remains uncertain, and what steps they should take.

These findings remain provisional until external review can be conducted. The study examines one incident and relies on public evidence, which cannot establish everything either organization communicated privately or did internally. Several criteria also draw on research from related fields. Further testing across incidents and reviewers is needed to assess the rubric's consistency, refine its scoring descriptions and examine how weighting affects the results.

Future work could expand the existing best practices into detailed guidance and action plans for each criterion. Each plan could explain what good practice looks like, how to address an identified weakness, who should take responsibility and what evidence would demonstrate improvement. For example, guidance on notifying affected parties could include steps for identifying recipients, drafting useful notices and following up as the investigation develops. Guidance on updates and corrections could provide templates for explaining changes and preserving earlier versions.

An interactive tool could connect assessment results to these action plans, helping organizations prioritize improvements and track their progress. The guidance could also be tested in simulated incidents involving several organizations, alongside communication professionals, technical teams and potential recipients. This would help assess both how easily the guidance can be used and how well it supports understanding and action.

Organizations should agree in advance who will communicate, how affected parties will be notified and how updates and corrections will be coordinated.<sup>29</sup> Developing the rubric in this

direction could make it useful both for reviewing past communication and for preparing teams to respond more clearly during future incidents.

## Notes

1. Alison N. Novak and M. Olguta Vilceanu, “‘The Internet Is Not Pleased’: Twitter and the 2017 Equifax Data Breach,” *The Communication Review* 22, no. 3 (2019): 196–221; Richard Knight and Jason R. C. Nurse, “A Framework for Effective Corporate Communication after Cyber Security Incidents,” *Computers & Security* 99 (2020): 102036; Yixin Zou, Shawn Danino, Kaiwen Sun, and Florian Schaub, “You ‘Might’ Be Affected: An Empirical Analysis of Readability and Usability Issues in Data Breach Notifications,” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York: ACM, 2019), paper 194.
2. OpenAI, “OpenAI and Hugging Face Partner to Address Security Incident during Model Evaluation,” July 21, 2026, updated July 28 and 29, 2026, <https://openai.com/index/hugging-face-model-evaluation-security-incident/>; OpenAI, “The Hugging Face Incident and the Road Ahead,” August 26, 2026, <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>; OpenAI, *OpenAI–Hugging Face Incident Technical Report* (San Francisco: OpenAI, August 26, 2026).
3. Hugging Face, “Security Incident Disclosure — July 2026,” Hugging Face Blog, July 16, 2026, <https://huggingface.co/blog/security-incident-july-2026>; OpenAI, “OpenAI and Hugging Face Partner”; Simon Willison, “Now We Have a Timeline of the OpenAI Accidental Attack against Hugging Face,” Simon Willison’s Weblog, August 7, 2026, <https://simonwillison.net/2026/Aug/7/openai-timeline/>.
4. W. Timothy Coombs, “Protecting Organization Reputations During a Crisis: The Development and Application of Situational Crisis Communication Theory,” *Corporate Reputation Review* 10, no. 3 (2007): 163–76; W. Timothy Coombs, “Situational Crisis Communication Theory (SCCT): Refining and Clarifying a Cognitive-Based Theory of Crisis Communication,” in *The Handbook of Crisis Communication*, 2nd ed., ed. W. Timothy Coombs and Sherry J. Holladay (Hoboken, NJ: Wiley-Blackwell, 2022), 193–204; An-Sofie Claeys, “The Benefits and Pitfalls of Stealing Thunder,” in Coombs and Holladay, *Handbook of Crisis Communication*, 99–111; Matthew W. Seeger, “Best Practices in Crisis Communication: An Expert Panel Process,” *Journal of Applied Communication Research* 34, no. 3 (2006): 232–44; World Health Organization, *Communicating Risk in Public Health Emergencies: A WHO Guideline for Emergency Risk Communication (ERC) Policy and Practice* (Geneva: World Health Organization, 2017); Centers for Disease Control and Prevention, *Crisis and Emergency Risk Communication (CERC): Introduction*, 2018 update (Atlanta: CDC, 2018).
5. Allen D. Householder, Garret Wassermann, Art Manion, and Chris King, *The CERT Guide to Coordinated Vulnerability Disclosure*, CMU/SEI-2017-SR-022 (Pittsburgh: Software Engineering Institute, Carnegie Mellon University, 2017); International Organization for Standardization and International Electrotechnical Commission, *ISO/IEC 29147:2018, Information Technology — Security Techniques — Vulnerability Disclosure*, 2nd ed. (Geneva: ISO/IEC, 2018); Forum of Incident Response and Security Teams (FIRST), *Guidelines and Practices for Multi-Party Vulnerability Coordination and Disclosure*, version 1.1 (FIRST, 2020).
6. OECD, *Towards a Common Reporting Framework for AI Incidents*, OECD Artificial Intelligence Papers, no. 34 (Paris: OECD Publishing, 2025), <https://doi.org/10.1787/f326d4ac-en>.
7. National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, NIST AI 600-1 (Gaithersburg, MD: NIST, 2024), <https://doi.org/10.6028/NIST.AI.600-1>; National Institute of Standards and Technology, *Artificial*

*Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1 (Gaithersburg, MD: NIST, 2023), MANAGE 4.3, <https://doi.org/10.6028/NIST.AI.100-1>.

8. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), OJ L, 12 July 2024, art. 55, para. 1, point c, and art. 73; European Commission, AI Office, *General-Purpose AI Code of Practice: Safety and Security Chapter*, 10 July 2025, Commitment 9.

9. Shayne Longpre et al., “FLARE-AI: Flaw Reporting for AI,” arXiv:2606.31567 (June 30, 2026), <https://arxiv.org/abs/2606.31567>; see also Shayne Longpre et al., “In-House Evaluation Is Not Enough: Towards Robust Third-Party Flaw Disclosure for General-Purpose AI,” arXiv:2503.16861 (2025); Sven Cattell, Avijit Ghosh, and Lucie-Aimée Kaffee, “Coordinated Flaw Disclosure for AI: Beyond Security Vulnerabilities,” arXiv:2402.07039 (2024).

10. Regulation (EU) 2024/1689, art. 73; European Commission, *General-Purpose AI Code of Practice*, Commitment 9; NIST, *Generative Artificial Intelligence Profile*.

11. OpenAI, “Hugging Face Incident and the Road Ahead”; OpenAI, *Incident Technical Report*; Ryan Greenblatt, Ajeya Cotra, and Hjalmar Wijk, “Brief Independent Investigation of Agents’ Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident,” METR, August 26, 2026, <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>.

12. Josh Hawley, “Chairman Hawley Launches Investigation into OpenAI for Hacking, Existential Risk of AI Products,” press release with letter to Sam Altman dated September 9, 2026, US Senate, September 10, 2026, <https://www.hawley.senate.gov/chairman-hawley-launches-investigation-into-openai-for-hacking-existential-risk-of-ai-products/>; Richard Blumenthal to Sam Altman, September 9, 2026, [https://www.blumenthal.senate.gov/imo/media/doc/20260909\\_-\\_openai\\_-\\_cot\\_and\\_rogue\\_agentspdf.pdf](https://www.blumenthal.senate.gov/imo/media/doc/20260909_-_openai_-_cot_and_rogue_agentspdf.pdf); Chris Van Hollen to Sam Altman, September 10, 2026, US Senate press release, <https://www.vanhollen.senate.gov/news/press-releases>.

13. Hugging Face, “Security Incident Disclosure.”

14. OpenAI, *Incident Technical Report*; Willison, “Now We Have a Timeline.”

15. OpenAI, “OpenAI and Hugging Face Partner.”

16. OpenAI, “OpenAI and Hugging Face Partner,” updates of July 28 and 29, 2026.

17. OpenAI, “Hugging Face Incident and the Road Ahead”; OpenAI, *Incident Technical Report*; Greenblatt, Cotra, and Wijk, “Brief Independent Investigation.”

18. Hawley, “Chairman Hawley Launches Investigation”; Blumenthal to Altman, September 9, 2026; Van Hollen to Altman, September 10, 2026; Associated Press, “Senators from Both Parties Question OpenAI on Breach of AI Startup Hugging Face,” September 10, 2026, <https://www.pbs.org/newshour/politics/senators-from-both-parties-question-openai-on-breach-of-ai-startup-hugging-face>.

19. OpenAI, “OpenAI and Hugging Face Partner”; OpenAI, *Incident Technical Report*.

20. OpenAI, “OpenAI and Hugging Face Partner”; OpenAI, *Incident Technical Report*; Hugo Larcher et al., “Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident,” Hugging Face Blog, July 27, 2026, revised through July 30, 2026, <https://huggingface.co/blog/agent-intrusion-technical-timeline>; revision history at <https://github.com/huggingface/blog/commits/main/agent-intrusion-technical-timeline.md>.
21. OpenAI, “OpenAI and Hugging Face Partner,” update of July 29, 2026; Greenblatt, Cotra, and Wijk, “Brief Independent Investigation”; Blumenthal to Altman, September 9, 2026.
22. Laura M. Arpan and David R. Roskos-Ewoldsen, “Stealing Thunder: Analysis of the Effects of Proactive Disclosure of Crisis Information,” *Public Relations Review* 31, no. 3 (2005): 425–33; An-Sofie Claeys and Verolien Cauberghe, “Crisis Response and Crisis Timing Strategies, Two Sides of the Same Coin,” *Public Relations Review* 38, no. 1 (2012): 83–88; Claeys, “Benefits and Pitfalls”; Thomas I. Vaughan-Johnston, Joshua J. Guyer, Katie L. Lawrence, and Leandre R. Fabrigar, “Disclosure and Identification Information Increase the Benefits of Stealing Thunder,” *Social Influence* 20, no. 1 (2025): 2447273; Yongtian Yu, Weiming Ye, and Kaihang Zhang, “Faster? Softer? Or More Formal? A Study on the Methods of Enterprises’ Crisis Response on Social Media,” *Mathematics* 13, no. 10 (2025): 1582; Yan Huang and Marcia DiStaso, “Responding to a Health Crisis on Facebook: The Effects of Response Timing and Message Appeal,” *Public Relations Review* 46, no. 3 (2020): 101909; Novak and Vilceanu, “Internet Is Not Pleased”; Coombs, “Protecting Organization Reputations.”
23. WHO, *Communicating Risk*; CDC, *CERC: Introduction*; Anne Marthe van der Bles, Sander van der Linden, Alexandra L. J. Freeman, and David J. Spiegelhalter, “The Effects of Communicating Uncertainty on Public Trust in Facts and Numbers,” *Proceedings of the National Academy of Sciences* 117, no. 14 (2020): 7672–83; Charlotte Dries, Michelle McDowell, Felix G. Rebitschek, and Christina Leuker, “When Evidence Changes: Communicating Uncertainty Protects against a Loss of Trust,” *Public Understanding of Science* 33 (2024), <https://doi.org/10.1177/09636625241228449>; Knight and Nurse, “Framework for Effective Corporate Communication”; Zou et al., “You ‘Might’ Be Affected”; Vaughan-Johnston et al., “Disclosure and Identification Information”; Federal Trade Commission, *In re Blackbaud, Inc.*, complaint (2024); US Securities and Exchange Commission, “SEC Charges Four Companies with Misleading Cyber Disclosures,” press release 2024-174, October 22, 2024.
24. OECD, *Defining AI Incidents and Related Terms*, OECD Artificial Intelligence Papers, no. 16 (Paris: OECD Publishing, 2024), <https://doi.org/10.1787/d1a8d965-en>; Regulation (EU) 2024/1689, art. 3, point 49; Adriana Placani, “Anthropomorphism in AI: Hype and Fallacy,” *AI and Ethics* 4 (2024): 691–98; Jem Bartholomew and Dhrumil Mehta, “How the Media Is Covering ChatGPT,” *Columbia Journalism Review*, May 26, 2023; Sanneke Kuipers and Marleen Schonheit, “Data Breaches and Effective Crisis Communication: A Comparative Analysis of Corporate Reputational Crises,” *Corporate Reputation Review* 25 (2022): 176–97; Carson Ezell, Xavier Roberts-Gaal, and Alan Chan, “Incident Analysis for AI Agents,” arXiv:2508.14231 (2025).
25. W. Timothy Coombs, *Reexamining Crisis Communication* (Gainesville, FL: Institute for Public Relations, 2026), <https://instituteforpr.org/reexamining-crisis-communication/>; Coombs, “Situational Crisis Communication Theory”; Kuipers and Schonheit, “Data Breaches and Effective Crisis Communication”; Soo-Yeon Kim and Jeong-Hyeon Lee, “How to Maximize the Effectiveness of Stealing Thunder in Crisis Communication: The Significance of Follow-Up Actions and Transparent Communication,” *Corporate Communications: An International Journal* 27, no. 3 (2022): 425–40; Wouter Jong, “Beyond the Snapshot: Rethinking Crisis Communication Theories in Dynamic Crisis Situations,” *Public Relations Review* 51 (2025): 102586; Peter H. Kim, Donald L. Ferrin, Cecily D. Cooper, and Kurt T. Dirks, “Removing the

Shadow of Suspicion: The Effects of Apology versus Denial for Repairing Competence- versus Integrity-Based Trust Violations,” *Journal of Applied Psychology* 89, no. 1 (2004): 104–18; Arthur W. Page Society, “The Page Principles,” accessed September 14, 2026, <https://page.org/site/the-page-principles>.

26. Coombs, “Protecting Organization Reputations”; Zou et al., “You ‘Might’ Be Affected”; European Data Protection Board, *Guidelines 9/2022 on Personal Data Breach Notification under GDPR*, version 2.0 (March 28, 2023), para. 90; Substance Abuse and Mental Health Services Administration, *SAMHSA’s Concept of Trauma and Guidance for a Trauma-Informed Approach*, HHS Publication No. (SMA) 14-4884 (Rockville, MD: SAMHSA, 2014).

27. Associated Press, *The Associated Press Statement of News Values and Principles* (New York: Associated Press, 2026); BBC, *Editorial Guidelines*, section 3, “Accuracy” (London: BBC, 2019); Kuipers and Schonheit, “Data Breaches and Effective Crisis Communication”; WHO, *Communicating Risk*; FIRST, *Multi-Party Vulnerability Coordination and Disclosure*; Jong, “Beyond the Snapshot.”

28. ISO/IEC 29147:2018; Householder et al., *CERT Guide to Coordinated Vulnerability Disclosure*; Forum of Incident Response and Security Teams, *PSIRT Services Framework*, version 1.1 (FIRST, 2020); Longpre et al., “In-House Evaluation Is Not Enough”; Ezell, Roberts-Gaal, and Chan, “Incident Analysis for AI Agents”; Associated Press, *Statement of News Values*.

## Bibliography

Arpan, Laura M., and David R. Roskos-Ewoldsen. "Stealing Thunder: Analysis of the Effects of Proactive Disclosure of Crisis Information." *Public Relations Review* 31, no. 3 (2005): 425–33.

Arthur W. Page Society. "The Page Principles." Accessed September 14, 2026.  
<https://page.org/site/the-page-principles>.

Associated Press. "Senators from Both Parties Question OpenAI on Breach of AI Startup Hugging Face." September 10, 2026.  
<https://www.pbs.org/newshour/politics/senators-from-both-parties-question-openai-on-breach-of-ai-startup-hugging-face>.

Associated Press. *The Associated Press Statement of News Values and Principles*. New York: Associated Press, 2026.

Bartholomew, Jem, and Dhruvil Mehta. "How the Media Is Covering ChatGPT." *Columbia Journalism Review*, May 26, 2023.

BBC. *Editorial Guidelines*. Section 3, "Accuracy." London: BBC, 2019.

Blumenthal, Richard. Letter to Sam Altman, September 9, 2026.  
<https://www.blumenthal.senate.gov/imo/media/doc/20260909-openai-cot-and-rogue-agentspdf.pdf>.

Cattell, Sven, Avijit Ghosh, and Lucie-Aimée Kaffee. "Coordinated Flaw Disclosure for AI: Beyond Security Vulnerabilities." arXiv:2402.07039 (2024).

Centers for Disease Control and Prevention. *Crisis and Emergency Risk Communication (CERC): Introduction*. 2018 update. Atlanta: CDC, 2018.

Claeys, An-Sofie. "The Benefits and Pitfalls of Stealing Thunder." In *The Handbook of Crisis Communication*, 2nd ed., edited by W. Timothy Coombs and Sherry J. Holladay, 99–111. Hoboken, NJ: Wiley-Blackwell, 2022.

Claeys, An-Sofie, and Verolien Cauberghe. "Crisis Response and Crisis Timing Strategies, Two Sides of the Same Coin." *Public Relations Review* 38, no. 1 (2012): 83–88.

Coombs, W. Timothy. "Protecting Organization Reputations During a Crisis: The Development and Application of Situational Crisis Communication Theory." *Corporate Reputation Review* 10, no. 3 (2007): 163–76.

Coombs, W. Timothy. *Reexamining Crisis Communication*. Gainesville, FL: Institute for Public Relations, 2026. <https://instituteforpr.org/reexamining-crisis-communication/>.

Coombs, W. Timothy. "Situational Crisis Communication Theory (SCCT): Refining and Clarifying a Cognitive-Based Theory of Crisis Communication." In *The Handbook of Crisis Communication*, 2nd ed., edited by W. Timothy Coombs and Sherry J. Holladay, 193–204. Hoboken, NJ: Wiley-Blackwell, 2022.

Dries, Charlotte, Michelle McDowell, Felix G. Rebitschek, and Christina Leuker. "When Evidence Changes: Communicating Uncertainty Protects against a Loss of Trust." *Public Understanding of Science* 33 (2024). <https://doi.org/10.1177/09636625241228449>.

European Commission, AI Office. *General-Purpose AI Code of Practice: Safety and Security Chapter*. July 10, 2025.

European Data Protection Board. *Guidelines 9/2022 on Personal Data Breach Notification under GDPR*. Version 2.0. March 28, 2023.

European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). *Official Journal of the European Union* L, July 12, 2024.

Ezell, Carson, Xavier Roberts-Gaal, and Alan Chan. "Incident Analysis for AI Agents." arXiv:2508.14231 (2025).

Federal Trade Commission. *In re Blackbaud, Inc.* Complaint and Decision and Order. 2024.

Forum of Incident Response and Security Teams. *Guidelines and Practices for Multi-Party Vulnerability Coordination and Disclosure*. Version 1.1. FIRST, 2020.

Forum of Incident Response and Security Teams. *PSIRT Services Framework*. Version 1.1. FIRST, 2020.

Greenblatt, Ryan, Ajeya Cotra, and Hjalmar Wijk. "Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident." METR, August 26, 2026. <https://metr.org/blog/2026-08-26-openai-hugging-face-incident-investigation/>.

Hawley, Josh. "Chairman Hawley Launches Investigation into OpenAI for Hacking, Existential Risk of AI Products." Press release and letter to Sam Altman, September 9–10, 2026. <https://www.hawley.senate.gov/chairman-hawley-launches-investigation-into-openai-for-hacking-existential-risk-of-ai-products/>.

Householder, Allen D., Garret Wassermann, Art Manion, and Chris King. *The CERT Guide to Coordinated Vulnerability Disclosure*. CMU/SEI-2017-SR-022. Pittsburgh: Software Engineering Institute, Carnegie Mellon University, 2017.

Huang, Yan, and Marcia DiStaso. "Responding to a Health Crisis on Facebook: The Effects of Response Timing and Message Appeal." *Public Relations Review* 46, no. 3 (2020): 101909.

Hugging Face. "Security Incident Disclosure — July 2026." Hugging Face Blog, July 16, 2026. <https://huggingface.co/blog/security-incident-july-2026>.

International Organization for Standardization and International Electrotechnical Commission. *ISO/IEC 29147:2018, Information Technology — Security Techniques — Vulnerability Disclosure*. 2nd ed. Geneva: ISO/IEC, 2018.

Jong, Wouter. "Beyond the Snapshot: Rethinking Crisis Communication Theories in Dynamic Crisis Situations." *Public Relations Review* 51 (2025): 102586.

Kim, Peter H., Donald L. Ferrin, Cecily D. Cooper, and Kurt T. Dirks. "Removing the Shadow of Suspicion: The Effects of Apology versus Denial for Repairing Competence- versus Integrity-Based Trust Violations." *Journal of Applied Psychology* 89, no. 1 (2004): 104–18.

Kim, Soo-Yeon, and Jeong-Hyeon Lee. "How to Maximize the Effectiveness of Stealing Thunder in Crisis Communication: The Significance of Follow-Up Actions and Transparent Communication." *Corporate Communications: An International Journal* 27, no. 3 (2022): 425–40.

Knight, Richard, and Jason R. C. Nurse. "A Framework for Effective Corporate Communication after Cyber Security Incidents." *Computers & Security* 99 (2020): 102036.

Kuipers, Sanneke, and Marleen Schonheit. "Data Breaches and Effective Crisis Communication: A Comparative Analysis of Corporate Reputational Crises." *Corporate Reputation Review* 25 (2022): 176–97.

Larcher, Hugo, Adrien Carreira, Raphaël G., and Christophe Rannou. "Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident." Hugging Face Blog, July 27, 2026, revised through July 30, 2026. <https://huggingface.co/blog/agent-intrusion-technical-timeline>.

Longpre, Shayne, Elaine Zhu, Carson Ezell, Avijit Ghosh, Sean McGregor, et al. "FLARE-AI: Flaw Reporting for AI." arXiv:2606.31567 (2026). <https://arxiv.org/abs/2606.31567>.

Longpre, Shayne, Kevin Klyman, Ruth E. Appel, Sayash Kapoor, Rishi Bommasani, et al. "In-House Evaluation Is Not Enough: Towards Robust Third-Party Flaw Disclosure for General-Purpose AI." arXiv:2503.16861 (2025).

National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. Gaithersburg, MD: NIST, 2023. <https://doi.org/10.6028/NIST.AI.100-1>.

National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*. NIST AI 600-1. Gaithersburg, MD: NIST, 2024. <https://doi.org/10.6028/NIST.AI.600-1>.

Novak, Alison N., and M. Olguta Vilceanu. "'The Internet Is Not Pleased': Twitter and the 2017 Equifax Data Breach." *The Communication Review* 22, no. 3 (2019): 196–221.

OECD. *Defining AI Incidents and Related Terms*. OECD Artificial Intelligence Papers, no. 16. Paris: OECD Publishing, 2024. <https://doi.org/10.1787/d1a8d965-en>.

OECD. *Towards a Common Reporting Framework for AI Incidents*. OECD Artificial Intelligence Papers, no. 34. Paris: OECD Publishing, 2025. <https://doi.org/10.1787/f326d4ac-en>.

OpenAI. "The Hugging Face Incident and the Road Ahead." August 26, 2026. <https://openai.com/index/hugging-face-incident-and-the-road-ahead/>.

OpenAI. "OpenAI and Hugging Face Partner to Address Security Incident during Model Evaluation." July 21, 2026, updated July 28 and 29, 2026. <https://openai.com/index/hugging-face-model-evaluation-security-incident/>.

OpenAI. *OpenAI–Hugging Face Incident Technical Report*. San Francisco: OpenAI, August 26, 2026.

Placani, Adriana. "Anthropomorphism in AI: Hype and Fallacy." *AI and Ethics* 4 (2024): 691–98.

Seeger, Matthew W. "Best Practices in Crisis Communication: An Expert Panel Process." *Journal of Applied Communication Research* 34, no. 3 (2006): 232–44.

Substance Abuse and Mental Health Services Administration. *SAMHSA's Concept of Trauma and Guidance for a Trauma-Informed Approach*. HHS Publication No. (SMA) 14-4884. Rockville, MD: SAMHSA, 2014.

US Securities and Exchange Commission. "SEC Charges Four Companies with Misleading Cyber Disclosures." Press release 2024-174, October 22, 2024.

Van Hollen, Chris. Letter to Sam Altman, September 10, 2026. Press release, US Senate. <https://www.vanhollen.senate.gov/news/press-releases>.

van der Bles, Anne Marthe, Sander van der Linden, Alexandra L. J. Freeman, and David J. Spiegelhalter. "The Effects of Communicating Uncertainty on Public Trust in Facts and Numbers." *Proceedings of the National Academy of Sciences* 117, no. 14 (2020): 7672–83.

Vaughan-Johnston, Thomas I., Joshua J. Guyer, Katie L. Lawrence, and Leandre R. Fabrigar. "Disclosure and Identification Information Increase the Benefits of Stealing Thunder." *Social Influence* 20, no. 1 (2025): 2447273. <https://doi.org/10.1080/15534510.2024.2447273>.

Willison, Simon. "Now We Have a Timeline of the OpenAI Accidental Attack against Hugging Face." Simon Willison's Weblog, August 7, 2026. <https://simonwillison.net/2026/Aug/7/openai-timeline/>.

World Health Organization. *Communicating Risk in Public Health Emergencies: A WHO Guideline for Emergency Risk Communication (ERC) Policy and Practice*. Geneva: World Health Organization, 2017.

Yu, Yongtian, Weiming Ye, and Kaihang Zhang. "Faster? Softer? Or More Formal? A Study on the Methods of Enterprises' Crisis Response on Social Media." *Mathematics* 13, no. 10 (2025): 1582. <https://doi.org/10.3390/math13101582>.

Zou, Yixin, Shawn Danino, Kaiwen Sun, and Florian Schaub. "You 'Might' Be Affected: An Empirical Analysis of Readability and Usability Issues in Data Breach Notifications." In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, paper 194. New York: ACM, 2019.



# Appendices

## Appendix A

Appendix A presents the 15 criteria used to assess organizational communication during an AI incident. For each criterion, the table identifies its communication area and provides an open-ended question to guide evaluators' judgement.

These questions assess the quality and completeness of the observed communication rather than asking evaluators to make a binary yes-or-no determination.

C11, C13, C14 and C15 are conditional and are applied only when relevant to the incident.

| Communication Area | Criterion                         | What it assesses                                                                                                                                                                                               | Apples when |
|--------------------|-----------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------|
| Timing             | C1. Timing of public disclosure   | Relative to when it learned of the incident, was the organization the first credible public source, or, if another party disclosed first, did it publish before the first subsequent coverage of the incident? | Always      |
| Incident account   | C2. Known and unknown information | Did the organization separate established facts from open questions, explain why important facts remained unknown, and commit to a dated next update?                                                          | Always      |
| Incident account   | C3. Certainty and reassurance     | Were known facts stated clearly and genuine uncertainties bounded, and was each reassuring claim supported by what the organization had examined and what risk remained?                                       | Always      |
| Incident account   | C4. Scope and affected parties    | Could readers determine what was reached, how the incident occurred, how far it extended and for how long, and whether their own systems, data or credentials were affected?                                   | Always      |
| Incident account   | C5. Incident discovery            | Did the organization explain how and when it learned of the incident and, where an external party raised it, who that party was?                                                                               | Always      |
| Incident framing   | C6. Incident classification       | Did the organization use an accurate public label for the incident and state the definition,                                                                                                                   | Always      |

|                           |                                             |                                                                                                                                                                                                                            |                                                                                                                                                                                                                                |
|---------------------------|---------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
|                           |                                             | regulatory regime or severity framework supporting it?                                                                                                                                                                     |                                                                                                                                                                                                                                |
| Incident framing          | C7. Consistent description of the AI system | Was the AI system described in consistent, mechanism-based language across statements and spokespeople, distinguishing its actions from human decisions and indirect effects, without overstating or downplaying its role? | Always                                                                                                                                                                                                                         |
| Incident framing          | C8. AI-system conduct and authorization     | Did the organization explain which AI system acted, where and under whose authority it operated, what it was authorized to do, which safeguards were present or relaxed, and what evidence supported the account?          | Always                                                                                                                                                                                                                         |
| Responsibility and remedy | C9. Responsibility and remedies             | Did the organization state its own role without shifting it onto other parties, update that acknowledgement as certainty improved, and report corrective actions that were delivered and can be checked from outside?      | Always                                                                                                                                                                                                                         |
| Protection and reach      | C10. Protective guidance                    | Were people and organizations that may have been exposed given accessible, specific and prioritized guidance on what to do, alongside what the organization itself was doing?                                              | Always                                                                                                                                                                                                                         |
| Protection and reach      | C11. Direct notice to affected parties      | Did the organization notify identifiable affected parties directly, through appropriate channels not compromised by the incident, with evidence of that notification that others can check?                                | The public record describes identifiable parties' systems, data, credentials or accounts as affected. N/A only where the organization has stated, with the scope of what it examined, that no identifiable party was affected. |
| Updates and coordination  | C12. Updates and corrections                | Were promised updates delivered and dated, substantive changes and corrections clearly labelled, superseded statements linked to current information, and open questions addressed in a closing account?                   | Always                                                                                                                                                                                                                         |

|                          |                                         |                                                                                                                                                                                                                         |                                                                                                                                                                                                         |
|--------------------------|-----------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Updates and coordination | C13. Coordination between organizations | Did the organizations' public accounts jointly explain who did what, acknowledge unresolved attribution and represent each other's positions, without contradictions or shared omissions?                               | More than one organization is publicly involved                                                                                                                                                         |
| External scrutiny        | C14. External finders and reviewers     | Did the organization promptly acknowledge, credit and keep engaging with outside finders, answer press questions, and leave independent reviewers free to state their own access, limitations and independence?         | An outside party found the incident, assessed it independently, or asked for comment                                                                                                                    |
| Protection and reach     | C15. Support for people harmed          | Was communication with affected people sensitive and practically supportive, offering an accessible human contact, warnings about follow-on scams, and ways to remove, reverse or challenge harmful content or actions? | The incident exposed identifiable people's sensitive data, produced outputs about identifiable people, or took actions on someone's behalf. N/A where only organizational systems or data were reached. |

# Independent Assessment of AI Incident Communication

## OpenAI and Hugging Face | July to August 2026

**Your task.** Assess how each organization communicated during the incident. Evaluate the public communication record, not the quality of the underlying technical response. Please complete the assessment independently before discussing it with anyone else.

|                       |  |
|-----------------------|--|
| Reviewer name or code |  |
| Date completed        |  |

## How to complete the assessment

1. Read the timeline on the following pages. Use the linked source documents when you need more detail.
2. Assess OpenAI and Hugging Face separately across the full communication period, from the first public disclosure through 26 August 2026.
3. For every criterion, enter one score and cite at least one dated source. Add a short sentence explaining your choice.
4. Use N/A only when the criterion does not apply or the public record is genuinely insufficient. Mark which reason applies.
5. Do not calculate an overall score. The research team will calculate results after collecting the independent assessments.

## Scoring scale

| Score | Meaning                    |
|-------|----------------------------|
| 0     | Absent or actively harmful |

|     |                                                                 |
|-----|-----------------------------------------------------------------|
| 1   | Minimal or significantly deficient                              |
| 2   | Substantially present, with important gaps                      |
| 3   | Strong practice meeting the full evidence-based standard        |
| N/A | Not applicable or insufficient public evidence - identify which |

## Evidence sources

The timeline is a navigation aid, not a substitute for the underlying communications.

Open the sources when the summary does not contain enough evidence to score a criterion.

- S1** [Hugging Face security incident disclosure](#) (16 July 2026)
- S2** [TechCrunch reporting on the Hugging Face intrusion](#) (20 July 2026)
- S3** [OpenAI initial incident statement](#) (21 July 2026)
- S4** [Sam Altman, Clément Delangue](#) X posts and [Delangue's travel announcement](#) (21, 23 July 2026)
- S5** [Hugging Face technical timeline](#) (27 July 2026, subsequently revised)
- S6** [Hugging Face technical timeline revision history](#) (28-30 July 2026)
- S7** [Sam Altman's podcast interview on the incident](#) (28 July 2026)
- S8** [OpenAI incident account and road ahead](#) (26 August 2026)
- S9** [METR and Redwood independent investigation](#) (26 August 2026)

## Communication timeline

| Date and milestone                    | Event or communication                                                                                                                                                                                     | Source ID |
|---------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|
| 16 Jul<br>Hugging Face discloses      | Hugging Face announced an intrusion driven by autonomous AI agents, described the known impact, advised users to rotate tokens and review activity, and said affected parties would be contacted directly. | S1        |
| 20 Jul<br>First mainstream coverage   | TechCrunch published the first mainstream news report of the Hugging Face intrusion (05:39 PDT, 12:39 UTC).                                                                                                | S2        |
| 21 Jul<br>OpenAI confirms involvement | OpenAI publicly stated that its models, including GPT-5.6 Sol and a more capable pre-release model with reduced cyber refusals, drove the incident.                                                        | S3        |
| 21 Jul<br>Sam Altman posts on X       | Sam Altman (OpenAI CEO) shared OpenAI's findings and thanked Hugging Face for its partnership on the incident.                                                                                             | S4        |

|                                                                     |                                                                                                                                                                                                                                                                                                                                                                                               |        |
|---------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------|
| <b>21, 23 Jul</b><br><b>Clément Delangue</b><br><b>X posts</b>      | Clément Delangue (Hugging Face CEO) discussed cooperation with OpenAI and the ongoing investigation on 21 July, and announced travel to San Francisco on 23 July.                                                                                                                                                                                                                             | S4     |
| <b>27-30 Jul</b><br><b>Hugging Face</b><br><b>technical account</b> | Hugging Face published a detailed reconstruction covering the attack path, agent actions, affected systems, detection, containment and remediation.<br>The public article was subsequently revised.                                                                                                                                                                                           | S5, S6 |
| <b>28 Jul</b><br><b>Sam Altman's</b><br><b>podcast interview</b>    | Sam Altman discussed the incident on <i>Invest Like the Best</i> with Patrick O'Shaughnessy, saying one of OpenAI's unreleased models "cheated" on a test by chaining zero-day exploits to escape its sandbox and break into Hugging Face's systems.<br>He says OpenAI paused training and needs to secure its sandboxes, and that AI development may need to slow down so society can adapt. | S7     |
| <b>26 Aug</b><br><b>OpenAI report and</b><br><b>external review</b> | OpenAI published a detailed incident account and described warning signals, investigative findings and new safeguards. METR and Redwood published a separate investigation of the agents' behavior the same day.                                                                                                                                                                              | S8, S9 |

## OpenAI assessment

Part 1 of 2. Enter one score for each criterion. In the evidence column, cite a source ID and date.

Do not compare OpenAI's score with the other organization while rating.

**Quick key:** 0 absent or harmful | 1 minimal | 2 substantial with gaps | 3 strong practice | N/A not applicable or insufficient evidence

| Criterion                              | What it assesses                                                                                                                                                                                               | Score | Evidence | Brief Reason |
|----------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|----------|--------------|
| <b>C1. Timing of public disclosure</b> | Relative to when it learned of the incident, was the organization the first credible public source, or, if another party disclosed first, did it publish before the first subsequent coverage of the incident? |       |          |              |

|                                                    |                                                                                                                                                                                                                            |  |  |  |
|----------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--|--|--|
| <b>C2. Known and unknown information</b>           | Did the organization separate established facts from open questions, explain why important facts remained unknown, and commit to a dated next update?                                                                      |  |  |  |
| <b>C3. Certainty and reassurance</b>               | Were known facts stated clearly and genuine uncertainties bounded, and was each reassuring claim supported by what the organization had examined and what risk remained?                                                   |  |  |  |
| <b>C4. Scope and affected parties</b>              | Could readers determine what was reached, how the incident occurred, how far it extended and for how long, and whether their own systems, data or credentials were affected?                                               |  |  |  |
| <b>C5. Incident discovery</b>                      | Did the organization explain how and when it learned of the incident and, where an external party raised it, who that party was?                                                                                           |  |  |  |
| <b>C6. Incident classification</b>                 | Did the organization use an accurate public label for the incident and state the definition, regulatory regime or severity framework supporting it?                                                                        |  |  |  |
| <b>C7. Consistent description of the AI system</b> | Was the AI system described in consistent, mechanism-based language across statements and spokespeople, distinguishing its actions from human decisions and indirect effects, without overstating or downplaying its role? |  |  |  |
| <b>C8. AI-system conduct and authorization</b>     | Did the organization explain which AI system acted, where and under whose authority it operated, what it was authorized to do, which safeguards were present or relaxed, and what evidence supported the account?          |  |  |  |
| <b>C9. Responsibility and remedies</b>             | Did the organization state its own role without shifting it onto other parties, update that acknowledgement as certainty improved, and report corrective actions that were delivered and can be checked from outside?      |  |  |  |
| <b>C10. Protective guidance</b>                    | Were people and organizations that may have been exposed given accessible, specific and prioritized guidance on what to do, alongside what the organization itself was doing?                                              |  |  |  |

|                                                |                                                                                                                                                                                                                         |  |  |  |
|------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--|--|--|
| <b>C11. Direct notice to affected parties</b>  | Did the organization notify identifiable affected parties directly, through appropriate channels not compromised by the incident, with evidence of that notification that others can check?                             |  |  |  |
| <b>C12. Updates and corrections</b>            | Were promised updates delivered and dated, substantive changes and corrections clearly labelled, superseded statements linked to current information, and open questions addressed in a closing account?                |  |  |  |
| <b>C13. Coordination between organizations</b> | Did the organizations' public accounts jointly explain who did what, acknowledge unresolved attribution and represent each other's positions, without contradictions or shared omissions?                               |  |  |  |
| <b>C14. External finders and reviewers</b>     | Did the organization promptly acknowledge, credit and keep engaging with outside finders, answer press questions, and leave independent reviewers free to state their own access, limitations and independence?         |  |  |  |
| <b>C15. Support for people harmed</b>          | Was communication with affected people sensitive and practically supportive, offering an accessible human contact, warnings about follow-on scams, and ways to remove, reverse or challenge harmful content or actions? |  |  |  |

## Hugging Face assessment

Part 2 of 2. Enter one score for each criterion. In the final column, cite a source ID and date, then give a one-sentence reason.

Do not compare Hugging Face's score with the other organization while rating.

**Quick key:** 0 absent or harmful | 1 minimal | 2 substantial with gaps | 3 strong practice | N/A applicable or insufficient evidence

| Criterion                              | What it assesses                                                                                                                                                                                               | Score | Evidence | Brief Reason |
|----------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------|----------|--------------|
| <b>C1. Timing of public disclosure</b> | Relative to when it learned of the incident, was the organization the first credible public source, or, if another party disclosed first, did it publish before the first subsequent coverage of the incident? |       |          |              |

|                                                    |                                                                                                                                                                                                                            |  |  |  |
|----------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--|--|--|
| <b>C2. Known and unknown information</b>           | Did the organization separate established facts from open questions, explain why important facts remained unknown, and commit to a dated next update?                                                                      |  |  |  |
| <b>C3. Certainty and reassurance</b>               | Were known facts stated clearly and genuine uncertainties bounded, and was each reassuring claim supported by what the organization had examined and what risk remained?                                                   |  |  |  |
| <b>C4. Scope and affected parties</b>              | Could readers determine what was reached, how the incident occurred, how far it extended and for how long, and whether their own systems, data or credentials were affected?                                               |  |  |  |
| <b>C5. Incident discovery</b>                      | Did the organization explain how and when it learned of the incident and, where an external party raised it, who that party was?                                                                                           |  |  |  |
| <b>C6. Incident classification</b>                 | Did the organization use an accurate public label for the incident and state the definition, regulatory regime or severity framework supporting it?                                                                        |  |  |  |
| <b>C7. Consistent description of the AI system</b> | Was the AI system described in consistent, mechanism-based language across statements and spokespeople, distinguishing its actions from human decisions and indirect effects, without overstating or downplaying its role? |  |  |  |
| <b>C8. AI-system conduct and authorization</b>     | Did the organization explain which AI system acted, where and under whose authority it operated, what it was authorized to do, which safeguards were present or relaxed, and what evidence supported the account?          |  |  |  |
| <b>C9. Responsibility and remedies</b>             | Did the organization state its own role without shifting it onto other parties, update that acknowledgement as certainty improved, and report corrective actions that were delivered and can be checked from outside?      |  |  |  |

|                                                |                                                                                                                                                                                                                         |  |  |  |
|------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--|--|--|
| <b>C10. Protective guidance</b>                | Were people and organizations that may have been exposed given accessible, specific and prioritized guidance on what to do, alongside what the organization itself was doing?                                           |  |  |  |
| <b>C11. Direct notice to affected parties</b>  | Did the organization notify identifiable affected parties directly, through appropriate channels not compromised by the incident, with evidence of that notification that others can check?                             |  |  |  |
| <b>C12. Updates and corrections</b>            | Were promised updates delivered and dated, substantive changes and corrections clearly labelled, superseded statements linked to current information, and open questions addressed in a closing account?                |  |  |  |
| <b>C13. Coordination between organizations</b> | Did the organizations' public accounts jointly explain who did what, acknowledge unresolved attribution and represent each other's positions, without contradictions or shared omissions?                               |  |  |  |
| <b>C14. External finders and reviewers</b>     | Did the organization promptly acknowledge, credit and keep engaging with outside finders, answer press questions, and leave independent reviewers free to state their own access, limitations and independence?         |  |  |  |
| <b>C15. Support for people harmed</b>          | Was communication with affected people sensitive and practically supportive, offering an accessible human contact, warnings about follow-on scams, and ways to remove, reverse or challenge harmful content or actions? |  |  |  |

## Overall observations

After completing every criterion, briefly identify the most important communication strength and weakness for each organization. Do not revise earlier scores simply to make them match this overall impression.

### OpenAI

**Most important strength:**

**Most important weakness:**

### Hugging Face

**Most important strength:**

**Most important weakness:**

### Criteria that were difficult to score

**Criterion ID and reason:**

**Completion check**

- ☐ I assessed the two organizations separately.
- ☐ Every numerical score has a dated source and short justification.
- ☐ Every N/A identifies whether it means not applicable or insufficient public evidence.
- ☐ I completed the assessment before discussing my ratings with other reviewers.

## Appendix C. Incident Timeline

# Timeline of Hugging Face-OpenAI Incident

This timeline tracks the operational events and communications linked to the July 2026 Hugging Face incident. The developments include internal warnings, actions before the intrusion, joint coordination, organizational response after the breach, the changing public narrative and subsequent investigations related to the Hugging Face incident.

| Date and milestone                                                              | Event or Communication                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       | Source ID |
|---------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------|
| <b>27 Jun<br/>Internal warning<br/>&amp; Initial<br/>Response</b>               | OpenAI's monitoring system detected anomalous scanning across network ports in its research environment and their on-call response staff were informed.                                                                                                                                                                                                                                                                                                                                                      | S1        |
| <b>11–13 Jul<br/>Incident<br/>Escalation and<br/>Containment</b>                | On July 11 <sup>th</sup> , agents accessed information about the execution environment of a Hugging Face Dataset Server worker at 4:38 UTC and executed OS-level commands at 16:07 UTC. On 13th July, Hugging Face severed agents' access to its infrastructure.                                                                                                                                                                                                                                             | S2        |
| <b>16 Jul<br/>Hugging Face<br/>discloses</b>                                    | Hugging Face publicly announced an intrusion driven by autonomous AI-agents and known scope of impact. It also advised users to rotate tokens, review and report suspicious account activities and committed to contacting affected customers and partners directly as required.                                                                                                                                                                                                                             | S3        |
| <b>16/17 Jul<br/>Customer Impact<br/>Inquiry by<br/>OpenAI</b>                  | OpenAI contacted Hugging Face asking if its data had been affected by the incident because it was a Hugging Face customer.                                                                                                                                                                                                                                                                                                                                                                                   | S4        |
| <b>19–20 Jul<br/>Joint<br/>Investigation by<br/>Hugging Face<br/>and OpenAI</b> | On July 19 <sup>th</sup> , OpenAI monitoring systems flagged unusual identity-related API calls associated with internal research Artifactory roles and responders initiated the shutdown of active ExploitGym evaluations. On July 20 <sup>th</sup> , OpenAI found Hugging Face credentials in the files preserved during its investigation and contacted Hugging Face to inquire about them. Hugging Face responded that it had already rotated two of those credentials during its own incident response. | S5        |
| <b>20 Jul<br/>First mainstream<br/>coverage</b>                                 | TechCrunch published the first mainstream news report of the Hugging Face intrusion (05:39 PDT, 12:39 UTC).                                                                                                                                                                                                                                                                                                                                                                                                  | S6        |
| <b>21 Jul<br/>OpenAI Confirms<br/>Involvement</b>                               | OpenAI made its involvement in the Hugging Face intrusion public through an article published on its website and identified the incident as being driven by a combination of OpenAI models including GPT-5.6 Sol and an even more capable pre-release model, all with reduced cyber refusals for evaluation purposes.                                                                                                                                                                                        | S7        |
| <b>27–30 Jul<br/>Technical<br/>Timeline<br/>Published by<br/>Hugging Face</b>   | Hugging Face published its technical post on 28 July at 20:08 UTC which had detailed reconstruction of how the intrusion unfolded, how it was detected and contained what remediation steps were taken.                                                                                                                                                                                                                                                                                                      | S8        |
| <b>26 Aug<br/>Post-Incident<br/>Accountability &amp;<br/>External Review</b>    | OpenAI published its 38-page technical incident report, after collaborating with METR and Redwood Research for third-party assessment of its model behavior. METR and Redwood Research jointly published their investigation report of agent behavior the same day on METR's website.                                                                                                                                                                                                                        | S9        |

## Sources:

- S1** OpenAI, [OpenAI–Hugging Face Incident Technical Report](#) (26 August 2026), p. 8.
- S2** OpenAI, [Incident Technical Report](#), p. 35; Hugging Face, Technical Timeline (27 July 2026), “Day 5” and “How we intercepted and analyzed the attack.”
- S3** Hugging Face, [Security Incident Disclosure](#), opening paragraphs, “What happened” and “For our community.”
- S4** OpenAI, Incident Technical Report, pp. 14–15; Simon Willison, [Now We Have a Timeline of the OpenAI Accidental Attack against Hugging Face](#) (7 August 2026), July 16 and July 20 entries.
- S5** OpenAI, Incident Technical Report, pp. 4, 15, 37–38.
- S6** TechCrunch, [Hugging Face confirms breach affected internal datasets and credentials, urges users to take action](#) (20 July 2026).
- S7** OpenAI, [Initial Incident Statement](#) (21 July 2026); Incident Technical Report, p. 4.
- S8** Hugging Face, [Technical Timeline](#) (27 July 2026); [Repository revision history](#) (28–30 July 2026).
- S9** OpenAI, Incident Technical Report, pp. 4–5, 7–8, 25–31; METR and Redwood Research, [Independent Investigation of Agents’ Behavior](#) (26 August 2026)

## Appendix D. Limitations and Dual-Use Considerations

We acknowledge that there is a risk that the rubric and scoring framework may be used to coach PR professionals in order to maximise public sympathy and/or protect company reputation, rather than be used for the intended application of public disclosure and harm reduction.

## Appendix E. Source Register

| ID    | Source                                                                  | Type                   | Year      | Issuing body / venue                                                                                 |
|-------|-------------------------------------------------------------------------|------------------------|-----------|------------------------------------------------------------------------------------------------------|
| S1-01 | Coombs, "Protecting Organization Reputations During a Crisis" (SCCT)    | Scholarly              | 2007      | Corporate Reputation Review 10(3): 163–176                                                           |
| S1-02 | Ma & Zhan, meta-analysis of SCCT research                               | Scholarly              | 2016      | J. Public Relations Research 28(2)                                                                   |
| S1-03 | Coombs, SCCT chapter, Handbook of Crisis Communication (2nd ed.)        | Scholarly              | 2022      | Handbook of Crisis Communication, 2nd ed. (Coombs & Holladay, eds), Wiley-Blackwell: ch. 14, 193–204 |
| S1-04 | Coombs & Tachkova, moral outrage in SCCT                                | Scholarly              | 2023      | Management Communication Quarterly 37(4): 798–820                                                    |
| S1-05 | Coombs, "Reexamining Crisis Communication" (IPR white paper)            | Practitioner/scholarly | May 2026  | Institute for Public Relations                                                                       |
| S1-06 | Coombs, "State of Crisis Communication: Evidence and the Bleeding Edge" | Practitioner/press     | 2014      | Institute for Public Relations                                                                       |
| S1-07 | Coombs, "Crisis Management and Communications"                          | Practitioner/press     | 2007/2014 | Institute for Public Relations                                                                       |
| S1-08 | Benoit, "Image repair discourse and crisis communication"               | Scholarly              | 1997      | Public Relations Review 23(2): 177–186                                                               |

|       |                                                                                                                                    |           |                   |                                                                                             |
|-------|------------------------------------------------------------------------------------------------------------------------------------|-----------|-------------------|---------------------------------------------------------------------------------------------|
| S1-10 | Seeger, "Best Practices in Crisis Communication: An Expert Panel Process"                                                          | Scholarly | 2006              | J. Applied Communication Research 34(3): 232–244                                            |
| S1-11 | Arpan & Roskos-Ewoldsen, "Stealing thunder: Analysis of the effects of proactive disclosure of crisis information"                 | Scholarly | 2005              | Public Relations Review 31(3): 425–433                                                      |
| S1-12 | Claeys & Cauberghe, "Crisis response and crisis timing strategies, two sides of the same coin"                                     | Scholarly | 2012              | Public Relations Review 38(1)                                                               |
| S1-14 | Claeys, "The Benefits and Pitfalls of Stealing Thunder"                                                                            | Scholarly | 2022              | Handbook of Crisis Communication, 2nd ed. (Coombs & Holladay, eds), Wiley-Blackwell: 99–111 |
| S1-15 | Beldad, van Laar & Hegner, "Should the shady steal thunder?"                                                                       | Scholarly | 2018              | J. Contingencies & Crisis Management 26(1): 150–163                                         |
| S1-16 | Kim & Lee, stealing thunder follow-up actions                                                                                      | Scholarly | 2022              | Corporate Communications 27(3): 425–440                                                     |
| S1-17 | Lee & Fraustino, stealing thunder for legal violations                                                                             | Scholarly | 2023 (print 2025) | J. Marketing Communications 31(3): 263–283                                                  |
| S1-18 | Ndone, "Beyond Negative Emotions: The Effects of Positive Emotions and Stealing Thunder on Organizational Outcomes After a Crisis" | Scholarly | 2025              | J. Contingencies & Crisis Management 33(3): e70075                                          |

|       |                                                                                                                                     |           |                        |                                                |
|-------|-------------------------------------------------------------------------------------------------------------------------------------|-----------|------------------------|------------------------------------------------|
| S1-19 | Fennis & Stroebe, "Softening the Blow"                                                                                              | Scholarly | 2014                   | J. Business Ethics 120(1): 109–120             |
| S1-20 | Vaughan-Johnston, Guyer, Lawrence & Fabrigar, "Disclosure and identification information increase the benefits of stealing thunder" | Scholarly | 2025 (online Dec 2024) | Social Influence 20(1)                         |
| S1-21 | Jin, Liu & Austin, Social-Mediated Crisis Communication model                                                                       | Scholarly | 2014 (online 2011)     | Communication Research 41(1): 74–94            |
| S1-22 | Yu, Ye & Zhang, enterprise crisis response on social media (3.1M posts, 94 incidents)                                               | Scholarly | 2025                   | Mathematics (MDPI) 13(10): 1582                |
| S1-23 | Huang & DiStaso, "Responding to a Health Crisis on Facebook: The Effects of Response Timing and Message Appeal"                     | Scholarly | 2020                   | Public Relations Review 46(3): 101909          |
| S1-24 | Upadhyay & Upadhyay, bibliometric map of crisis comms 1968–2022                                                                     | Scholarly | 2023                   | Humanities & Social Sciences Communications 10 |
| S1-25 | Zhou, Zhang & Ki, "Crisis response strategy manipulation: A systematic review and a test of nuances"                                | Scholarly | 2022                   | Public Relations Review 48: 102208             |
| S1-26 | Jong, "Beyond the snapshot: Rethinking crisis communication theories in dynamic crisis situations"                                  | Scholarly | 2025                   | Public Relations Review 51: 102586             |

|       |                                                                                                                                            |           |                    |                                            |
|-------|--------------------------------------------------------------------------------------------------------------------------------------------|-----------|--------------------|--------------------------------------------|
| S1-27 | Macnamara, insider perspective during COVID-19                                                                                             | Scholarly | 2021               | Public Relations Inquiry 10(2): 237–262    |
| S1-28 | Knight & Nurse, "A framework for effective corporate communication after cyber security incidents"                                         | Scholarly | 2020               | Computers & Security                       |
| S1-29 | Kuipers & Schonheit, "Data Breaches and Effective Crisis Communication: A Comparative Analysis of Corporate Reputational Crises" (8 cases) | Scholarly | 2022 (online 2021) | Corporate Reputation Review 25(3): 176–197 |
| S1-30 | Novak & Vilceanu, Twitter and the 2017 Equifax breach                                                                                      | Scholarly | 2019               | The Communication Review                   |
| S1-31 | Bentley & Ma, "Testing perceptions of organizational apologies after a data breach crisis"                                                 | Scholarly | 2020               | Public Relations Review 46(5): 101975      |
| S1-32 | Ruuhonen, Hjerpe & Kortesoja, "Crisis Communication in the Face of Data Breaches" (Vastaamo)                                               | Scholarly | 2024               | arXiv 2406.01744                           |
| S1-33 | Jong, Mikelatou & Vielvoije, "When crisis communication meets forensics" (11 practitioner interviews)                                      | Scholarly | 2026               | Computers & Security 171: 105082           |

|       |                                                                                                            |                   |             |                                              |
|-------|------------------------------------------------------------------------------------------------------------|-------------------|-------------|----------------------------------------------|
| S1-34 | Xiao, Zhou & Yu, "Unraveling the anchoring effect of crisis communication in cyberattack spillover crises" | Scholarly         | 2024        | Public Relations Review 50(2): 102449        |
| S1-35 | Coombs & Holladay, "Comparing apology to equivalent crisis response strategies"                            | Scholarly         | 2008        | Public Relations Review 34(3)                |
| S1-37 | CDC, Crisis and Emergency Risk Communication (CERC) Manual, Introduction                                   | Standard/guidance | 2018 update | US CDC                                       |
| S1-39 | WHO, "Communicating risk in public health emergencies" (GRADE guideline)                                   | Standard/guidance | 2017        | World Health Organization                    |
| S1-40 | Jha et al., systematic review for the 2017 WHO emergency risk communication guideline                      | Scholarly         | 2018        | PLOS ONE 13(10): e0205555                    |
| S1-41 | UK Government Communication Service, Crisis communications operating model                                 | Standard/guidance | 2023        | GCS                                          |
| S1-42 | GCS, STOP crisis communications planning guide                                                             | Standard/guidance | 24 Sep 2024 | GCS                                          |
| S1-43 | AIDR, Public Information and Warnings Handbook                                                             | Standard/guidance | 2021        | Australian Institute for Disaster Resilience |
| S1-44 | Australian Government Crisis Management Framework, v4.2                                                    | Standard/guidance | Sep 2026    | PM&C                                         |

|       |                                                                                   |                    |                       |                                                       |
|-------|-----------------------------------------------------------------------------------|--------------------|-----------------------|-------------------------------------------------------|
| S1-45 | Ewart & Heinrichs, communication gaps in Australian disasters (186 inquiries)     | Scholarly          | 2025                  | Risk, Hazards & Crisis in Public Policy 16(2): e70013 |
| S1-46 | CISA et al., "Communicating Under Pressure: Best Practices for Service Providers" | Standard/guidance  | 2 Sep 2026            | CISA and partners                                     |
| S1-47 | PRSA Code of Ethics                                                               | Policy/code        | 2000                  | PRSA                                                  |
| S1-48 | CIPR Code of Conduct                                                              | Policy/code        | current               | CIPR                                                  |
| S1-50 | IPRA Code of Conduct                                                              | Policy/code        | 2011                  | IPRA                                                  |
| S1-51 | Arthur W. Page Society, Page Principles                                           | Policy/code        | current               | Page Society                                          |
| S1-52 | Global Alliance Global Code of Ethics                                             | Policy/code        | 2018, re-adopted 2026 | Global Alliance for PR                                |
| S1-55 | Sandman, "Risk = Hazard + Outrage"                                                | Practitioner/press | 1987– (e.g. ENR 1999) | Peter Sandman                                         |
| S1-56 | Slovic, "Perceived Risk, Trust, and Democracy"                                    | Scholarly          | 1993                  | Risk Analysis 13(6): 675–682                          |
| S1-59 | Siegrist, "Trust and Risk Perception: A Critical Review"                          | Scholarly          | 2021 (online 2019)    | Risk Analysis 41(3): 480–490                          |
| S1-61 | Fischhoff & Davis, "Communicating scientific uncertainty"                         | Scholarly          | 2014                  | PNAS                                                  |
| S1-62 | van der Bles et al., effects of communicating uncertainty on trust                | Scholarly          | 2020                  | PNAS                                                  |

|       |                                                                                                    |           |      |                                                          |
|-------|----------------------------------------------------------------------------------------------------|-----------|------|----------------------------------------------------------|
| S1-63 | Kerr et al., uncertainty around statistics and public trust (N = 10,519)                           | Scholarly | 2023 | Royal Society Open Science 10(11): 230604                |
| S1-64 | Kreps & Kriner, model uncertainty and trust in science (COVID)                                     | Scholarly | 2020 | Science Advances 6(43): eabd4563                         |
| S1-65 | Dries et al., "When evidence changes: communicating uncertainty protects against a loss of trust"  | Scholarly | 2024 | Public Understanding of Science 33(6): 777–794           |
| S1-66 | Kim, Ferrin, Cooper & Dirks, apology vs denial for competence vs integrity violations              | Scholarly | 2004 | J. Applied Psychology                                    |
| S1-67 | Lewicki, Polin & Lount, structure of effective apologies                                           | Scholarly | 2016 | Negotiation & Conflict Management Research 9(2): 177–196 |
| S1-68 | Sharma, Schoorman & Ballinger, "How Can It Be Made Right Again? A Review of Trust Repair Research" | Scholarly | 2023 | J. Management 49(1): 363–399                             |
| S1-69 | Ho & Liu, apology laws and malpractice                                                             | Scholarly | 2011 | J. Risk and Uncertainty 43(2): 141–167                   |
| S1-70 | McMichael, Van Horn & Viscusi, "'Sorry' Is Never Enough"                                           | Scholarly | 2019 | Stanford Law Review                                      |
| S1-71 | Arbel-Groissman et al., apologies and likelihood of lawsuits                                       | Scholarly | 2025 | J. Medical Internet Research 27: e77493                  |

|       |                                                              |                   |            |                                     |
|-------|--------------------------------------------------------------|-------------------|------------|-------------------------------------|
| S2-01 | Guidelines 9/2022 on personal data breach notification, v2.0 | Standard/guidance | 2023       | EDPB                                |
| S2-02 | Provvedimento n. 755 (OpenAI/ChatGPT)                        | Case/enforcement  | 2 Nov 2024 | Garante (IT)                        |
| S2-03 | ICO, "Personal data breaches: a guide"                       | Standard/guidance | 2025       | UK ICO                              |
| S2-04 | Sullivan & Maniff, "Data Breach Notification Laws"           | Scholarly         | 2016       | Federal Reserve Bank of Kansas City |
| S2-05 | Kesari, "Do Data Breach Notification Laws Work?"             | Scholarly         | 2023–24    | NYU J. Legislation & Public Policy  |
| S2-06 | GDPR Arts 33–34 and Recitals 85–88                           | Legislation       | 2016       | EU                                  |
| S2-07 | CISA Coordinated Vulnerability Disclosure Process            | Standard/guidance | current    | CISA                                |
| S2-08 | Binding Operational Directive 20-01                          | Legislation       | 2020       | CISA/DHS                            |
| S2-09 | CERT/CC Vulnerability Disclosure Policy                      | Policy/code       | current    | CERT/CC                             |
| S2-10 | CERT Guide to Coordinated Vulnerability Disclosure           | Standard/guidance | 2017/web   | CERT/CC (SEI)                       |
| S2-11 | FIRST PSIRT Services Framework v1.1                          | Standard/guidance | 2020       | FIRST                               |
| S2-12 | FIRST Multi-Party Vulnerability Coordination guidelines v1.1 | Standard/guidance | 2020       | FIRST                               |
| S2-13 | Google Project Zero, Policy and Disclosure 2021              | Policy/code       | 2021       | Google                              |

|       |                                                                                          |                    |          |                               |
|-------|------------------------------------------------------------------------------------------|--------------------|----------|-------------------------------|
| S2-14 | Project Zero "Reporting Transparency" trial                                              | Policy/code        | Jul 2025 | Google                        |
| S2-15 | Microsoft Coordinated Vulnerability Disclosure                                           | Policy/code        | current  | Microsoft                     |
| S2-16 | Zero Day Initiative disclosure policy                                                    | Policy/code        | current  | Trend Micro ZDI               |
| S2-17 | CNA Operational Rules v4.1.0                                                             | Standard/guidance  | May 2025 | CVE Program                   |
| S2-18 | ENISA, Coordinated Vulnerability Disclosure Policies in the EU                           | Standard/scholarly | 2022     | ENISA                         |
| S2-19 | NIS2 Directive Arts 12, 23; Annex I                                                      | Legislation        | 2022     | EU                            |
| S2-20 | Implementing Regulation 2024/2690 (significant-incident thresholds)                      | Legislation        | 2024     | European Commission           |
| S2-21 | NIS2 transposition: Commission press release IP/26/1499 (8 Jul 2026) and status trackers | Practitioner/press | 2026     | European Commission; trackers |
| S2-22 | DORA Art. 19 and RTS 2025/301                                                            | Legislation        | 2025     | EU                            |
| S2-23 | DOJ Justice Manual 9-48.000 (CFAA charging policy)                                       | Standard/guidance  | 2022     | US DOJ                        |
| S2-24 | disclose.io safe-harbour framework                                                       | Policy/code        | current  | disclose.io                   |
| S2-25 | HackerOne Disclosure Guidelines                                                          | Policy/code        | current  | HackerOne                     |
| S2-26 | Bugcrowd Standard Disclosure Terms                                                       | Policy/code        | current  | Bugcrowd                      |

|       |                                                                                              |                   |                                                                                            |                     |
|-------|----------------------------------------------------------------------------------------------|-------------------|--------------------------------------------------------------------------------------------|---------------------|
| S2-27 | ISO/IEC 29147:2018<br>Vulnerability disclosure (2nd ed.)                                     | Standard/guidance | 2018                                                                                       | ISO/IEC JTC 1/SC 27 |
| S2-29 | SEC final rule, Cybersecurity Risk Management, Strategy, Governance, and Incident Disclosure | Legislation       | 2023                                                                                       | US SEC              |
| S2-30 | Gerding statement on Item 1.05 vs 8.01                                                       | Standard/guidance | May 2024                                                                                   | SEC Corp Fin        |
| S2-31 | SEC settled orders: Unisys, Avaya, Check Point, Mimecast                                     | Case/enforcement  | Oct 2024                                                                                   | US SEC              |
| S2-32 | SEC v. SolarWinds complaint                                                                  | Case/enforcement  | Oct 2023                                                                                   | US SEC              |
| S2-33 | SEC v. SolarWinds, opinion on motion to dismiss (Engelmayer J.), 107 pp.                     | Case/enforcement  | 18 Jul 2024                                                                                | S.D.N.Y.            |
| S2-34 | SEC v. SolarWinds voluntary dismissal (LR-26423)                                             | Case/enforcement  | Nov 2025                                                                                   | US SEC              |
| S2-35 | Security of Critical Infrastructure Act 2018 Part 2B                                         | Legislation       | 2022                                                                                       | Australia           |
| S2-36 | Cyber Security Act 2024 s 27 and Cyber Security (Ransomware Payment Reporting) Rules 2025    | Legislation       | Rules made 27 Feb 2025; commencement tied to Part 3 of the Act (30 May 2025 per secondary) | Australia           |
| S2-37 | ASX Guidance Note 8, Annexure A Example I (data breach)                                      | Standard/guidance | 2024                                                                                       | ASX                 |

|       |                                                            |                    |                 |                     |
|-------|------------------------------------------------------------|--------------------|-----------------|---------------------|
| S2-38 | HIPAA Breach Notification Rule, 45 CFR 164.404–.414        | Legislation        | 2013            | HHS OCR             |
| S2-39 | California Civil Code §1798.82 (as amended by SB 446)      | Legislation        | eff. 1 Jan 2026 | California          |
| S2-40 | FTC Health Breach Notification Rule, 16 CFR 318            | Legislation        | 2024            | US FTC              |
| S2-41 | GLBA Safeguards Rule 16 CFR 314.4(j)                       | Legislation        | 2024            | US FTC              |
| S2-42 | In re Blackbaud (FTC complaint, joint statement, order)    | Case/enforcement   | 2024            | US FTC              |
| S2-43 | In re Uber Technologies (FTC revised complaint)            | Case/enforcement   | 2018            | US FTC              |
| S2-44 | Privacy Act 1988 (Cth) Part IIIC, Notifiable Data Breaches | Legislation        | 2018            | Australia           |
| S2-45 | OAIC Notifiable Data Breaches Report, Jul–Dec 2024         | Practitioner/press | 2025            | OAIC                |
| S2-46 | Privacy and Other Legislation Amendment Act 2024           | Legislation        | 2024            | Australia           |
| S2-47 | AIC v Medibank Private                                     | Case/enforcement   | filed 2024      | OAIC; Federal Court |
| S2-48 | AIC v Singtel Optus                                        | Case/enforcement   | filed Aug 2025  | OAIC; Federal Court |

|        |                                                                                                                                                                                                                                 |                   |            |                                  |
|--------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------|------------|----------------------------------|
| S2-49  | Australian Information Commissioner v Australian Clinical Labs Limited (No 2) [2025] FCA 1224 (Halley J)                                                                                                                        | Case/enforcement  | 8 Oct 2025 | Federal Court of Australia       |
| S2-50  | In re Capital One Customer Data Security Breach Litigation (E.D. Va., Mag. J. Anderson, 26 May 2020); Wengui v. Clark Hill (D.D.C., Boasberg J., 12 Jan 2021); In re Rutter's (M.D. Pa., Chief Mag. J. Mehalchick, 22 Jul 2021) | Case/enforcement  | 2020–21    | US district courts               |
| S2B-01 | GPAI Code of Practice, Safety & Security chapter, Commitment 9                                                                                                                                                                  | Standard/guidance | Jul 2025   | EU AI Office; independent chairs |
| S2B-02 | OECD, "Towards a Common Reporting Framework for AI Incidents"                                                                                                                                                                   | Standard/guidance | Feb 2025   | OECD/GPAI                        |
| S2B-03 | NIST AI 600-1 Generative AI Profile                                                                                                                                                                                             | Standard/guidance | Jul 2024   | NIST                             |
| S2B-04 | NIST AI RMF 1.0                                                                                                                                                                                                                 | Standard/guidance | Jan 2023   | NIST                             |
| S2B-05 | OpenAI Preparedness Framework v2                                                                                                                                                                                                | Policy/code       | Apr 2025   | OpenAI                           |
| S2B-06 | Anthropic Responsible Scaling Policy v3.0                                                                                                                                                                                       | Policy/code       | Feb 2026   | Anthropic                        |
| S2B-07 | Google DeepMind Frontier Safety Framework v3.0                                                                                                                                                                                  | Policy/code       | 2025       | Google DeepMind                  |
| S2B-08 | America's AI Action Plan                                                                                                                                                                                                        | Standard/guidance | Jul 2025   | White House OSTP                 |

|        |                                                                                                            |                    |                      |                     |
|--------|------------------------------------------------------------------------------------------------------------|--------------------|----------------------|---------------------|
| S2B-09 | CSET, "AI Incidents: Key Components for a Mandatory Reporting Regime"                                      | Scholarly          | Jan 2025             | CSET Georgetown     |
| S2B-10 | Humane Intelligence GRT Transparency Report                                                                | Practitioner/press | 2024                 | Humane Intelligence |
| S2B-11 | GPAI Code Vademecum, Signatory Taskforce                                                                   | Standard/guidance  | Jan 2026             | EU AI Office        |
| S2B-12 | EU AI Act Art. 3(49) serious incident definition                                                           | Legislation        | 2024                 | EU                  |
| S2B-13 | EU AI Act Art. 73                                                                                          | Legislation        | 2024                 | EU                  |
| S2B-14 | EU AI Act Art. 55(1)(c)                                                                                    | Legislation        | 2024                 | EU                  |
| S2B-15 | Commission draft guidance and template on Art. 73 serious incidents                                        | Standard/guidance  | Sep 2025 (draft)     | European Commission |
| S2B-16 | Latham & Watkins alert on draft Art. 73 guidance                                                           | Practitioner/press | Oct 2025             | Latham & Watkins    |
| S2B-17 | California SB 53, Transparency in Frontier AI Act                                                          | Legislation        | operative 1 Jan 2026 | California          |
| S2B-18 | New York RAISE Act, S6953-B, Chapter 699 of 2025 (signed 19 Dec 2025); reported chapter amendment Mar 2026 | Legislation        | 2025–26              | New York            |
| S2B-19 | Colorado SB 24-205 → SB 25B-004 → SB 26-189                                                                | Legislation        | eff. 1 Jan 2027      | Colorado            |

|        |                                                                                                                                                                                                                            |                           |          |                            |
|--------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------|----------|----------------------------|
| S2B-20 | Australia: Voluntary AI Safety Standard (Sep 2024), superseded in practice by Guidance for AI Adoption (6 practices, 21 Oct 2025); mandatory guardrails shelved Dec 2025; "Australian Standards for AI" announced Jul 2026 | Standard/guidance         | 2024–26  | DISR; PM&C                 |
| S2B-21 | EO 14148 rescinding EO 14110                                                                                                                                                                                               | Legislation               | Jan 2025 | White House                |
| S2B-22 | NIST AI 100-2 E2025 Adversarial ML taxonomy                                                                                                                                                                                | Standard/guidance         | Mar 2025 | NIST                       |
| S2B-23 | CAISI outputs 2025–26                                                                                                                                                                                                      | Standard/guidance         | 2025–26  | NIST CAISI                 |
| S2B-24 | OECD, "Defining AI incidents and related terms"                                                                                                                                                                            | Standard/guidance         | May 2024 | OECD                       |
| S2B-25 | OECD AI Incidents and Hazards Monitor methodology                                                                                                                                                                          | Practitioner/presentation | 2024–    | OECD.AI                    |
| S2B-26 | G7 Hiroshima AI Process Reporting Framework                                                                                                                                                                                | Standard/guidance         | Feb 2025 | OECD for G7                |
| S2B-27 | Frontier AI Safety Commitments (Seoul)                                                                                                                                                                                     | Policy/code               | May 2024 | UK/ROK; 20 companies       |
| S2B-28 | International AI Safety Report 2026                                                                                                                                                                                        | Scholarly                 | Feb 2026 | International expert panel |
| S2B-29 | Longpre et al., "A Safe Harbor for AI Evaluation and Red Teaming"                                                                                                                                                          | Scholarly                 | 2024     | arXiv 2403.04893; ICML     |

|        |                                                                                                 |                   |             |                        |
|--------|-------------------------------------------------------------------------------------------------|-------------------|-------------|------------------------|
| S2B-30 | Longpre, Klyman, Appel, Kapoor, Bommasani et al., "In-House Evaluation Is Not Enough"           | Scholarly         | 2025        | arXiv 2503.16861       |
| S2B-31 | Cattell, Ghosh & Kaffee, "Coordinated Flaw Disclosure for AI"                                   | Scholarly         | 2024        | arXiv 2402.07039; AIES |
| S2B-32 | Kolt et al., "Responsible Reporting for Frontier AI Development"                                | Scholarly         | 2024        | arXiv 2404.02675       |
| S2B-33 | Longpre, Zhu, Ezell, Ghosh, McGregor et al., "FLARE-AI: Flaw Reporting for AI"                  | Scholarly         | Jun 2026    | arXiv 2606.31567; ICML |
| S2B-34 | Ezell, Roberts-Gaal & Chan, "Incident Analysis for AI Agents"                                   | Scholarly         | Aug 2025    | arXiv 2508.14231       |
| S2B-35 | Abraham et al., "AI Incident Monitoring through a Public Health Lens"                           | Scholarly         | Apr 2026    | arXiv 2604.19914       |
| S2B-36 | Stanford AI Index 2026, Responsible AI chapter                                                  | Scholarly         | 2026        | Stanford HAI           |
| S2B-37 | CVE AI Working Group Charter v1.1; CNA Rules §4.2.3                                             | Standard/guidance | 2024–25     | CVE Program            |
| S2B-38 | CVE Program blog, "CVE ID Assignment and CVE Record Publication for AI-Related Vulnerabilities" | Standard/guidance | 18 Feb 2025 | CVE Program            |

|        |                                                                                   |                    |          |                                 |
|--------|-----------------------------------------------------------------------------------|--------------------|----------|---------------------------------|
| S2B-39 | OpenAI CVE Assignment Policy (20 Nov 2025); Bug Bounty announcement (11 Apr 2023) | Policy/code        | 2023–25  | OpenAI                          |
| S2B-40 | Google AI Vulnerability Reward Program rules                                      | Policy/code        | Oct 2025 | Google                          |
| S2B-41 | Meta Bug Bounty GenAI payout guidelines                                           | Policy/code        | 2024–26  | Meta                            |
| S2B-42 | Anthropic model-safety bug bounty                                                 | Policy/code        | May 2025 | Anthropic                       |
| S2B-43 | Hacking Policy Council analysis of Copyright Office §1201 ruling                  | Practitioner/press | Dec 2024 | HPC                             |
| S2B-44 | DEF CON AI Village Generative Red Team 2 and 3                                    | Practitioner/press | 2024–25  | AI Village et al.               |
| S2B-45 | MITRE ATLAS AI Incident Sharing initiative                                        | Practitioner/press | Oct 2024 | MITRE CTID                      |
| S2B-46 | CLTR, "AI incident reporting: addressing a gap in UK regulation"                  | Practitioner/press | Jun 2024 | Centre for Long-Term Resilience |
| S2B-47 | CLTR, "Preparing for AI security incidents"                                       | Practitioner/press | Sep 2025 | CLTR                            |
| S2B-48 | AI Incident Database, Editor's Guide                                              | Practitioner/press | current  | Responsible AI Collaborative    |
| S2B-49 | AIAAIC Repository                                                                 | Practitioner/press | current  | AIAAIC                          |
| S2B-50 | AVID, AI Vulnerability Database                                                   | Practitioner/press | current  | AVID                            |

|        |                                                                                       |                        |          |                                 |
|--------|---------------------------------------------------------------------------------------|------------------------|----------|---------------------------------|
| S2B-51 | MIT AI Risk Repository v4 and Incident Tracker                                        | Practitioner/p<br>ress | Dec 2025 | MIT FutureTech                  |
| S2B-52 | OWASP Top 10 for LLM Applications 2025                                                | Standard/gui<br>dance  | 2025     | OWASP GenAI Security Project    |
| S2B-53 | GDPR Art. 4(12) personal data breach                                                  | Legislation            | 2016     | EU                              |
| S2B-54 | 6 U.S.C. §650 "incident" (CIRCA baseline)                                             | Legislation            | 2022     | US                              |
| S2B-55 | Hugging Face Hub security documentation                                               | Policy/code            | current  | Hugging Face                    |
| S3-01  | Zou, Danino, Sun & Schaub, "You 'Might' Be Affected" (161 notices)                    | Scholarly              | 2019     | CHI 2019                        |
| S3-02  | Zou, Mysore & Schaub, "I've Got Nothing to Lose" (Equifax interviews)                 | Scholarly              | 2018     | SOUPS 2018                      |
| S3-03  | Mayer, Zou, Schaub & Aviv, "'Now I'm a bit angry'" (n = 413, real breaches)           | Scholarly              | 2021     | USENIX Security 2021            |
| S3-04  | Ablon et al., Consumer Attitudes Toward Data Breach Notifications (RAND, n = 2,038)   | Scholarly              | 2016     | RAND                            |
| S3-05  | Ponemon/Experian, "The Aftermath of a Data Breach: Consumer Sentiment"                | Practitioner/p<br>ress | Apr 2014 | Ponemon (Experian-sponsored)    |
| S3-06  | Romanosky, Telang & Acquisti, "Do Data Breach Disclosure Laws Reduce Identity Theft?" | Scholarly              | 2011     | J. Policy Analysis & Management |

|       |                                                                                                                      |                        |            |                                                   |
|-------|----------------------------------------------------------------------------------------------------------------------|------------------------|------------|---------------------------------------------------|
| S3-09 | Bisogni, "Proving Limits of State Data Breach Notification Laws" (445 notices)                                       | Scholarly              | 2016       | J. Information Policy 6: 154–205                  |
| S3-10 | Shen et al., "An incident may have resulted in a suspected data compromise" (n = 99)                                 | Scholarly              | 2022       | SOUPS 2022 poster                                 |
| S3-11 | Bruemmer (Experian), "The misconceptions of data breach fatigue"                                                     | Practitioner/p<br>ress | 2016       | IAPP                                              |
| S3-12 | Shaver, "Breach Fatigue: Consumer Apathy Towards Data Breaches and Personal Security" (undergraduate honours thesis) | Scholarly              | May 2016   | Ohio State University, Fisher College of Business |
| S3-13 | OAIC, Australian Community Attitudes to Privacy Survey 2023 (n = 1,916)                                              | Practitioner/p<br>ress | 2023       | OAIC                                              |
| S3-14 | Morgan & Voce, "Data breaches and cybercrime victimisation"                                                          | Scholarly              | 2022       | Australian Institute of Criminology               |
| S3-17 | TermScout, "What's Market? A look into Data Breach Notifications"                                                    | Practitioner/p<br>ress | 9 Aug 2021 | TermScout                                         |
| S3-18 | FTC, "Data Breach Response: A Guide for Business" and model letter                                                   | Standard/gui<br>dance  | 2016       | FTC                                               |
| S3-22 | OAIC, Data breach preparation and response, Parts 3–4                                                                | Standard/gui<br>dance  | current    | OAIC                                              |

|       |                                                                                                                                                                          |                        |                         |                                |
|-------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------|-------------------------|--------------------------------|
| S3-29 | Vastaamo breach: Finnish Ministry of Social Affairs and Health victim-support page (Oct 2020) and Yle (Feb 2026); crisis-comms analysis via Ruohonen et al. 2024 (S1-32) | Primary/pres<br>s      | 2020–26                 | Finnish Government; Yle        |
| S3-30 | Ashley Madison FTC settlement                                                                                                                                            | Case/enforc<br>ement   | 2016                    | FTC                            |
| S3-31 | Scamwatch (ACCC), Optus data breach scams alert                                                                                                                          | Standard/gui<br>dance  | 2022                    | ACCC                           |
| S3-32 | Medibank newsroom: customer updates and support package                                                                                                                  | Primary<br>record      | 2022                    | Medibank                       |
| S3-33 | 2022 Optus breach SMS blackmail case: AFP sentencing release (7 Feb 2023) and ABC (6 Oct 2022)                                                                           | Primary/pres<br>s      | 2022–23                 | Australian Federal Police; ABC |
| S3-34 | KnowBe4 Q3 2017 phishing subject lines (via Infosecurity)                                                                                                                | Practitioner/p<br>ress | 2017                    | KnowBe4                        |
| S3-35 | Bîzgă, fake breach notifications; Kantra, "Scammers Are Now Sending Fake 'Your Data Was Breached' Emails"                                                                | Practitioner/p<br>ress | Jun<br>2024/Mar<br>2026 | Bitdefender; Techlicious       |
| S3-36 | Wash, Nthala & Rader, non-expert phishing detection (n = 297)                                                                                                            | Scholarly              | 2021                    | SOUPS 2021                     |

|       |                                                                                       |                      |            |                                |
|-------|---------------------------------------------------------------------------------------|----------------------|------------|--------------------------------|
| S3-37 | SAMHSA, Concept of Trauma and Guidance for a Trauma-Informed Approach                 | Standard/guidance    | 2014       | SAMHSA                         |
| S3-39 | eSafety Commissioner, image-based abuse and deepfakes guidance                        | Standard/guidance    | 2024–25    | eSafety (AU)                   |
| S3-40 | StopNCII.org, "How it works"                                                          | Practitioner/press   | current    | SWGfL                          |
| S3-41 | Cyber Civil Rights Initiative Safety Center                                           | Practitioner/press   | current    | CCRI                           |
| S3-42 | TAKE IT DOWN Act (Pub. L. 119-12, 19 May 2025) and FTC enforcement blog (19 May 2026) | Legislation/guidance | 2025–26    | Congress; FTC                  |
| S3-43 | CDT, NCII hash-matching governance brief                                              | Practitioner/press   | 9 Apr 2026 | CDT                            |
| S3-44 | ITRC 2025 Consumer Impact Report                                                      | Practitioner/press   | Oct 2025   | Identity Theft Resource Center |
| S3-45 | IDCARE                                                                                | Practitioner/press   | current    | IDCARE                         |
| S3-46 | ISO 24495-1:2023 Plain language, Part 1                                               | Standard/guidance    | 2023       | ISO TC37                       |
| S3-47 | Plain Writing Act 2010; Federal Plain Language Guidelines and testing guidance        | Legislation/guidance | 2010/2011  | US GSA; PLAIN                  |
| S3-48 | Australian Government Style Manual, "Literacy and access"                             | Standard/guidance    | 2024       | DTA                            |

|       |                                                                                   |                   |           |                                                          |
|-------|-----------------------------------------------------------------------------------|-------------------|-----------|----------------------------------------------------------|
| S3-49 | European Commission, "How to write clearly"                                       | Standard/guidance | 2015      | European Commission DG Translation (Publications Office) |
| S3-50 | Redish, limitations of readability formulas                                       | Scholarly         | 2000      | ACM J. Computer Documentation 24(3): 132–137             |
| S3-51 | Garcia v. Character Technologies; Character.AI "Community Safety Updates"         | Case/primary      | 2024–26   | M.D. Fla.; Character.AI                                  |
| S3-52 | Moffatt v. Air Canada, 2024 BCCRT 149                                             | Case/enforcement  | 2024      | BC Civil Resolution Tribunal                             |
| S3-53 | Carlini et al., "Extracting Training Data from Large Language Models"             | Scholarly         | 2021      | USENIX Security 2021                                     |
| S3-55 | OpenAI, "How the voices for ChatGPT were chosen"; Johansson statement coverage    | Primary/pres      | May 2024  | OpenAI; NPR/TechCrunch                                   |
| S3-56 | FTC consumer alert on AI-enhanced family emergency scams; Voice Cloning Challenge | Standard/guidance | 2023–24   | FTC                                                      |
| S3-57 | EU AI Act Art. 86 right to explanation; Commission timeline page                  | Legislation       | 2024/2026 | EU                                                       |
| S3-58 | GDPR Art. 22                                                                      | Legislation       | 2016      | EU                                                       |
| S3-59 | Kaminski & Urban, "The Right to Contest AI"                                       | Scholarly         | 2021      | Columbia Law Review                                      |

|       |                                                                                                                            |                        |                               |                                     |
|-------|----------------------------------------------------------------------------------------------------------------------------|------------------------|-------------------------------|-------------------------------------|
| S3-60 | Contestability literature (Alfrink et al. 2022/23; Lyons, Velloso & Miller 2021; Yurrita et al. 2025; Stewart, FAccT 2026) | Scholarly              | 2021–26                       | ACM/arXiv                           |
| S3-61 | CRA Art. 14(8) user-notification duty (lead-verified)                                                                      | Legislation            | 2024<br>(applies 11 Sep 2026) | EU                                  |
| S4-01 | SPJ Code of Ethics                                                                                                         | Policy/code            | 2014                          | Society of Professional Journalists |
| S4-02 | AP Statement of News Values and Principles                                                                                 | Policy/code            | 2026 ed.                      | Associated Press                    |
| S4-03 | Reuters Standards and Values (Handbook of Journalism)                                                                      | Policy/code            | current                       | Reuters                             |
| S4-04 | BBC Editorial Guidelines, Section 3 Accuracy                                                                               | Policy/code            | current                       | BBC                                 |
| S4-05 | NYT Ethical Journalism handbook                                                                                            | Policy/code            | 2025                          | New York Times                      |
| S4-06 | Guardian News & Media Editorial Code                                                                                       | Policy/code            | Mar 2026                      | GNM                                 |
| S4-07 | MEAA Journalist Code of Ethics                                                                                             | Policy/code            | 1999 rev.                     | MEAA                                |
| S4-08 | Australian Press Council Statement of General Principles                                                                   | Policy/code            | 2014 (page current)           | Australian Press Council            |
| S4-09 | Zacharia & Grotto, "How to Responsibly Report on Hacks and Disinformation"                                                 | Practitioner/p<br>ress | 2020                          | Stanford Cyber Policy Center        |
| S4-10 | Allsop, "Should the media report on hacked campaign documents?"                                                            | Practitioner/p<br>ress | 2024                          | Columbia Journalism Review          |

|       |                                                                                            |                    |         |                                        |
|-------|--------------------------------------------------------------------------------------------|--------------------|---------|----------------------------------------|
| S4-11 | EFF Coders' Rights, Vulnerability Reporting FAQ                                            | Practitioner/press | current | EFF                                    |
| S4-12 | Science Media Centre, "10 best practice guidelines for reporting science & health stories" | Practitioner/press | 2012    | SMC UK                                 |
| S4-13 | SMC UK "For journalists" resources                                                         | Practitioner/press | current | SMC UK                                 |
| S4-14 | Australian Science Media Centre / Scimex                                                   | Practitioner/press | current | AusSMC                                 |
| S4-15 | SciLine                                                                                    | Practitioner/press | current | AAAS                                   |
| S4-16 | Boykoff & Boykoff, "Balance as bias"                                                       | Scholarly          | 2004    | Global Environmental Change            |
| S4-17 | Dixon & Clarke, "Heightening Uncertainty Around Certain Science"                           | Scholarly          | 2013    | Science Communication                  |
| S4-18 | Jensen, hedging and credibility in cancer news (N = 601)                                   | Scholarly          | 2008    | Human Communication Research           |
| S4-19 | Peters & Dunwoody, scientific uncertainty in media (special issue)                         | Scholarly          | 2016    | Public Understanding of Science        |
| S4-20 | Fleerackers et al., preprints in COVID digital media                                       | Scholarly          | 2022    | Health Communication                   |
| S4-21 | KSJ Science Editing Handbook v2.1                                                          | Practitioner/press | 2020–   | Knight Science Journalism @ MIT        |
| S4-22 | AHCJ Statement of Principles                                                               | Policy/code        | current | Association of Health Care Journalists |

|       |                                                                                       |                    |            |                            |
|-------|---------------------------------------------------------------------------------------|--------------------|------------|----------------------------|
| S4-23 | ASAPbio preprint guidance for journalists                                             | Practitioner/press | c. 2020    | ASAPbio                    |
| S4-24 | Verification Handbook (2014) and for Disinformation and Media Manipulation (2020)     | Practitioner/press | 2014/2020  | European Journalism Centre |
| S4-25 | First Draft, "Verifying online information" (five pillars)                            | Practitioner/press | 2020       | First Draft                |
| S4-26 | Bellingcat resources and editorial standards                                          | Practitioner/press | current    | Bellingcat                 |
| S4-27 | Hunt, "Here's how I verify data breaches"                                             | Practitioner/press | 2016       | Troy Hunt                  |
| S4-28 | Hunt, "Data breach disclosure 101"                                                    | Practitioner/press | 2017       | Troy Hunt                  |
| S4-29 | Schwartz, "Ransomware Groups' Data Leak Blogs Lie"                                    | Practitioner/press | 2024       | ISMG; GovInfoSecurity      |
| S4-30 | DataBreaches.net, ransomware listing confusion                                        | Practitioner/press | 2025       | DataBreaches.net           |
| S4-31 | Bitdefender, "Claimed Twice"                                                          | Practitioner/press | Jun 2026   | Bitdefender                |
| S4-32 | Recent Breaches editorial standards                                                   | Practitioner/press | current    | GalaxyWarden (commercial)  |
| S4-33 | Whittaker, "A data breach at analytics giant Mixpanel leaves a lot of open questions" | Practitioner/press | 2 Dec 2025 | TechCrunch                 |
| S4-34 | Thorne, "We read 100 data breach notifications to make this guide"                    | Practitioner/press | 2017       | Atlassian Statuspage       |

|       |                                                                            |                        |      |                                             |
|-------|----------------------------------------------------------------------------|------------------------|------|---------------------------------------------|
| S4-35 | Brennen, Howard & Nielsen, "An Industry-Led Debate"                        | Scholarly              | 2018 | Reuters Institute                           |
| S4-36 | Nielsen, "How news coverage, often uncritical, helps build up the AI hype" | Scholarly              | 2024 | Reuters Institute                           |
| S4-37 | Chuan, Tsai & Cho, framing AI in US newspapers                             | Scholarly              | 2019 | AIES 2019                                   |
| S4-38 | Ouchchy, Coin & Dubljević, AI ethics in the headlines                      | Scholarly              | 2020 | AI & Society                                |
| S4-39 | Nguyen & Hekman, news framing of AI (N = 3,098)                            | Scholarly              | 2022 | AI & Society                                |
| S4-40 | Roe & Perkins, UK ChatGPT headlines (671)                                  | Scholarly              | 2023 | Humanities & Social Sciences Communications |
| S4-41 | Bartholomew & Mehta, "How the media is covering ChatGPT"                   | Practitioner/p<br>ress | 2023 | CJR; Tow Center                             |
| S4-42 | Placani, "Anthropomorphism in AI: hype and fallacy"                        | Scholarly              | 2024 | AI and Ethics                               |
| S4-43 | Vicsek, Feher & Nguyen, framing AI in US elite media                       | Scholarly              | 2026 | Global Media and Communication              |
| S4-44 | Dodds et al., "AI Hype in Journalism" (special issue)                      | Scholarly              | 2026 | Digital Journalism                          |
| S4-45 | Simon, Nielsen & Fletcher, Generative AI and News Report 2025              | Scholarly              | 2025 | Reuters Institute                           |
| S4-46 | Reuters Institute Digital News Report 2025                                 | Practitioner/p<br>ress | 2025 | Reuters Institute                           |

|       |                                                                                             |                    |            |                                      |
|-------|---------------------------------------------------------------------------------------------|--------------------|------------|--------------------------------------|
| S4-47 | Axios Codebook, "The myth of the sophisticated hacker"                                      | Practitioner/press | 2019       | Axios                                |
| S4-48 | Carson & Singh, "Debunking the 5 Myths of Sophisticated Cyber Attacks"                      | Practitioner/press | 6 Apr 2017 | Security Magazine                    |
| S4-50 | Jenkins, Anandarajan & D'Ovidio, impression management in breach notification (212 letters) | Scholarly          | 2014       | Western J. Communication             |
| S4-51 | Veltsos, breach notifications as negative news                                              | Scholarly          | 2012       | Business Communication Quarterly     |
| S4-52 | Bentley, Oostman & Shah, apologies in ambiguous crises                                      | Scholarly          | 2018       | J. Contingencies & Crisis Management |
| S4-55 | Kim, Johnson & Park, five retail breaches and press framing                                 | Scholarly          | 2017       | Cogent Business & Management         |
| S4-56 | Equifax incident-handling case study                                                        | Scholarly          | 2018       | Issues in Information Systems        |
| S4-58 | Fuoli & Paradis, model of trust-repair discourse                                            | Scholarly          | 2014       | J. Pragmatics                        |
| S4-59 | Fuoli & Hart, trust-building strategies experiment                                          | Scholarly          | 2018       | Discourse & Society 29(5): 514–552   |
| S4-62 | Wiz Research, exposed DeepSeek database                                                     | Practitioner/press | Jan 2025   | Wiz                                  |
| S4-63 | Serrapere, "Data Breach Notifications Are Too Complex"                                      | Practitioner/press | 2020       | Chicago Policy Review                |

|       |                                                                 |                        |          |                    |
|-------|-----------------------------------------------------------------|------------------------|----------|--------------------|
| S4-64 | GIJN and Pulitzer Center, AI Accountability Reporting Guide     | Practitioner/p<br>ress | Jun 2026 | GIJN               |
| S4-65 | GIJN, Guide to Investigating Social Media Algorithms            | Practitioner/p<br>ress | 2025     | GIJN               |
| S4-66 | GIJN, Reporter's Guide to Detecting AI-Generated Content        | Practitioner/p<br>ress | 2024–25  | GIJN               |
| S4-67 | GIJN, Reporter's Guide to Investigating Digital Threats         | Practitioner/p<br>ress | c. 2023  | GIJN               |
| S4-68 | The Markup, "Show Your Work"                                    | Practitioner/p<br>ress | 2020–    | The Markup         |
| S4-69 | AlgorithmWatch publications                                     | Practitioner/p<br>ress | 2019–    | AlgorithmWatch     |
| S4-72 | Partnership on AI, Responsible Practices for Synthetic Media    | Standard/gui<br>dance  | 2023     | Partnership on AI  |
| S4-73 | JournalismAI research reports                                   | Practitioner/p<br>ress | 2019–25  | LSE Polis          |
| S4-74 | Poynter AI Ethics Guidelines / Starter Kit                      | Practitioner/p<br>ress | 2024–25  | Poynter            |
| S4-75 | Walkley Foundation AI and Journalism training                   | Practitioner/p<br>ress | 2023–24  | Walkley Foundation |
| S4-76 | American Journalism Project, AI for Local Reporting field guide | Practitioner/p<br>ress | 2025     | AJP                |
| S4-78 | Pulitzer Center AI Accountability Network / AI Spotlight        | Practitioner/p<br>ress | 2022–    | Pulitzer Center    |

|        |                                                                                                                |                    |                    |                      |
|--------|----------------------------------------------------------------------------------------------------------------|--------------------|--------------------|----------------------|
| S5O-01 | OpenAI, "March 20 ChatGPT outage: Here's what happened" (Wayback captures 24 Mar, 20 Apr, 21 Apr 2023)         | Primary record     | 2023               | OpenAI               |
| S5O-03 | Altman (@sama), X post acknowledging the ChatGPT chat-history bug (20:16 UTC)                                  | Primary record     | 22 Mar 2023        | X                    |
| S5O-04 | @JordanLWheeler, X post reporting other users' chat titles visible in ChatGPT (16:26 UTC)                      | Primary record     | 20 Mar 2023        | X                    |
| S5O-09 | Help Net Security, email wording                                                                               | Practitioner/press | 27 Mar 2023        | Help Net Security    |
| S5O-12 | Hacker News thread 35291112                                                                                    | Primary record     | 24 Mar 2023        | Hacker News          |
| S5O-13 | Garante temporary limitation order and press release                                                           | Case/enforcement   | 30–31 Mar 2023     | Garante              |
| S5O-15 | Garante press release                                                                                          | Case/enforcement   | 20 Dec 2024        | Garante              |
| S5O-16 | Tribunale di Roma annulment of the Garante OpenAI fine: ANSA (20 Mar 2026) and Il Sole 24 Ore NT+ (3 Jun 2026) | Practitioner/press | Mar–Jun 2026       | ANSA; Il Sole 24 Ore |
| S5O-17 | OpenAI, "What to know about a recent Mixpanel security incident" (Wayback 27 Nov, 11 Dec 2025; 27 Jan 2026)    | Primary record     | 2025–26            | OpenAI               |
| S5O-18 | Taylor, "Our response to a recent security incident"                                                           | Primary record     | Nov 2025 (undated) | Mixpanel             |

|                     |                                                                                          |                    |                |                            |
|---------------------|------------------------------------------------------------------------------------------|--------------------|----------------|----------------------------|
| S5O-19              | Hacker News post with full OpenAI email text                                             | Primary record     | 27 Nov 2025    | Hacker News                |
| S5O-26 /27          | OpenAI, "Sycophancy in GPT-4o" and "Expanding on what we missed with sycophancy"         | Primary record     | Apr–May 2025   | OpenAI                     |
| S5O-28              | Altman (@sama) and OpenAI, X posts on the GPT-4o sycophancy update                       | Primary record     | 25–30 Apr 2025 | X                          |
| S5O-32 (S5H-2 6/27) | OpenAI, Hugging Face incident posts (21 Jul 2026, updates 28–29 Jul; 26 Aug 2026 report) | Primary record     | 2026           | OpenAI                     |
| S5H-01 /02          | Hugging Face, "Space secrets security update" (raw GitHub; Wayback)                      | Primary record     | 31 May 2024    | Hugging Face               |
| S5H-03              | GitHub API commit history, huggingface/blog                                              | Primary record     | 2023–26        | GitHub                     |
| S5H-05              | Wiz Research, Hugging Face infrastructure risks                                          | Practitioner/press | 4 Apr 2024     | Wiz                        |
| S5H-06              | Wiggers, Hugging Face unauthorized access                                                | Practitioner/press | 31 May 2024    | TechCrunch                 |
| S5H-11              | The Register, Lasso token exposure, with Delangue statement                              | Practitioner/press | Dec 2023       | The Register               |
| S5H-12              | Hacker News thread 40544875                                                              | Primary record     | 1 Jun 2024     | Hacker News                |
| S5H-13 /14          | Hugging Face forum topics 89696, 68768; skops PR #421                                    | Primary record     | 2024           | Hugging Face forum; GitHub |
| S5H-16              | Hugging Face blog, Wiz partnership post                                                  | Primary record     | 4 Apr 2024     | Hugging Face               |

|           |                                                                                                        |                    |                |                              |
|-----------|--------------------------------------------------------------------------------------------------------|--------------------|----------------|------------------------------|
| S5H-18/19 | Hugging Face, "Security incident disclosure — July 2026" (raw) and PR #3465 metadata                   | Primary record     | 15–16 Jul 2026 | Hugging Face; GitHub         |
| S5H-21/22 | Hugging Face, "Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline" and its nine commits   | Primary record     | 27–30 Jul 2026 | Hugging Face                 |
| S5H-23/24 | Whittaker; BleepingComputer                                                                            | Practitioner/press | 20 Jul 2026    | TechCrunch; BleepingComputer |
| S5H-25    | Ha, Delangue "radical transparency"                                                                    | Practitioner/press | 26 Jul 2026    | TechCrunch                   |
| S5H-28    | METR and Redwood Research independent investigation                                                    | Scholarly/primary  | 26 Aug 2026    | METR                         |
| S5H-31    | Willison, OpenAI timeline post                                                                         | Practitioner/press | 7 Aug 2026     | Simon Willison               |
| S5H-32    | Politico, "OpenAI models shared hacking tips on secret messaging board before Hugging Face breach"     | Practitioner/press | 5 Aug 2026     | Politico                     |
| S5H-36/37 | Senate letters to OpenAI: Hawley (9 Sep 2026; release 10 Sep), Blumenthal (9 Sep), Van Hollen (10 Sep) | Primary record     | Sep 2026       | US Senate                    |
| S5H-40    | status.huggingface.co incident 965809                                                                  | Primary record     | 16 Jul 2026    | Hugging Face                 |
| S5H-43    | Truffle Security, 787 live Hugging Face tokens in training data                                        | Practitioner/press | Jun 2026       | Truffle Security             |

|              |                                                                                                                                                   |                    |                |                                                            |
|--------------|---------------------------------------------------------------------------------------------------------------------------------------------------|--------------------|----------------|------------------------------------------------------------|
| S5H-44/46/48 | Lasso Security blog (Dec 2023); BleepingComputer on JFrog (Feb 2024); ReversingLabs "nullifAI" (Feb 2025)                                         | Practitioner/press | 2023–25        | Lasso Security; BleepingComputer; ReversingLabs            |
| S5H-47       | JFrog blog, malicious Hugging Face models                                                                                                         | Practitioner/press | Feb 2024       | JFrog                                                      |
| S5H-49       | Altman (@sama), X post on the Hugging Face incident, linking the OpenAI statement (20:13 UTC)                                                     | Primary record     | 21 Jul 2026    | X                                                          |
| S5H-50       | Delangue (@ClementDelangue), X posts responding to OpenAI attribution (21 Jul, 20:50 UTC) and announcing a San Francisco trip (23 Jul, 11:04 UTC) | Primary record     | 21–23 Jul 2026 | X                                                          |
| S5H-51       | Altman, "How to Make an Abundant Future" (interview), Invest Like the Best EP.484 (Hugging Face incident discussion 16:14–17:44 on Spotify)       | Primary record     | 28 Jul 2026    | Invest Like the Best with Patrick O'Shaughnessy (Colossus) |