
OpenAI/Hugging Face AI incident is a cybercrime

Karabo Lerato Mokoena
Michalsons and the Cape Institute for Safe AI

With
Apart Research

Abstract

AI is being entrenched in all aspects of our societies, so the framing that it is so novel that our current laws do not apply and that lawmaking is too slow to respond is both dangerous and false. This framing risks treating AI companies and the AI models they build as operating in a legal vacuum. It allows AI companies to “move fast” and break existing laws without legal accountability. I challenge that premise by examining whether existing legal frameworks can govern AI incidents. To do this, I applied South African laws to the OpenAI/Hugging Face AI incident. I map the agents’ documented conduct onto the offences in Chapter 2 of the Cybercrimes Act 19 of 2020, test attribution to OpenAI under the Electronic Communications and Transactions Act 25 of 2002 and section 332(1) of the Criminal Procedure Act 51 of 1977, and set out the notification duty under section 22 of POPIA. I found that the existing legal system does apply, but only where the agent has a known owner. The legal framework should undergo targeted amendments to strengthen its application, especially for autonomous agents with unknown owners. I also note that South Africa’s combination of statute, customary law, common law principles of legal and constitutional interpretation provides some flexibility to apply existing law to emerging technologies, while a more rigid system may not be able to do the same. AI companies and the models they build do not exist outside the law. By enforcing existing laws, we can deter AI risks by incentivising AI companies to innovate safely, securely, and responsibly.

Introduction

AI is everything, everywhere and all at once. While this technology can lower barriers and help nations, individuals, and companies achieve the outcomes they seek more effectively, anything left untamed, like a bush left to grow wild, will eventually consume everything around it. Luckily for us, the law applies to almost everything and everyone equitably, or at least it should. If the law is so all-encompassing, why do AI companies behave as if it does not apply to them and their AI models? I cannot answer that question, but I can assure you the law does currently address the AI risk introduced by AI models. This project argues that existing laws apply to the OpenAI cyberattack on Hugging Face, and OpenAI can be held liable for cybercrimes under South African law.

Why regulation does not necessarily stifle innovation

Reframing how regulation affects innovation is important to enabling a safe, secure digital future.¹ Regulation can promote innovation. According to Porter, well-crafted regulation can drive innovation and offset its costs.² This is because companies do not have all the insight into all possible profitable, innovative avenues. Accordingly, Porter hypothesised that regulation can help companies overcome organisational inertia by reducing complacency and exposing firms to new issues, possibilities, and technologies. Porter's hypothesis holds in some settings. For example, environmental laws such as the CO2 emissions standards for cars created new markets for electric vehicles. In effect, car manufacturers have expanded EV production, contributed to lower emissions, increased mobility, and driven innovations in battery efficiency.

Environmental regulations and emissions standards created profitable avenues for companies in EV technologies while mitigating the known risks of climate change. On the effect of regulation on AI innovation, I agree with Porter that well-crafted AI regulations can promote innovation and offset its cost, specifically as AI companies struggle to offset the costs of AI advancements.

Enforce existing laws while we figure the really hard stuff out

Shoshana Zuboff puts it plainly.³ AI companies know the law applies to them and their AI models. They also know that some of their behaviour does not comply with the law (they pirated copyrighted material to train their models).⁴ That is why they lobby, issue conflicting and emotional doomsday statements about AI capabilities, and try to explain away the legal misalignment of their AI models with complex AI jargon that makes it difficult to connect the act with a legal trigger. But a hack is a hack, right? Unauthorised access or theft of data is not novel. Humans and other capable tools have done this for years and faced the consequences. Existing laws apply even where an AI agent is the threat actor.

Let's find ways to address gaps in current laws

For now, we can still tie a human to an AI agent's actions. Noam Kolt has already discussed how Agency law applies between a company and its AI agent.⁵ An administrative small-claims tribunal in British Columbia can provide some perspective. In *Moffatt v Air Canada*, the airline was held liable for a negligent misrepresentation made by its chatbot.⁶ This supports the argument that OpenAI can be held liable for its agents' actions against Hugging Face. Governing AI agents becomes harder when we don't know who owns them. Kristie

¹ Mokoena, K. L. (2026, May). Chapter 7: Porter's hypothesis: Reframing the effects of the GDPR on AI innovation. University of Groningen Press. Legal Challenges in Data Driven Innovation. <https://opentextbooks.rug.nl/currentlegalchallengesindatadriveninnovation/chapter/chapter-7-porters-hypothesis-reframing-the-effects-of-the-gdpr-on-ai-innovation/>.

² Porter, M., & van der Linde, C. (1995). Toward a new conception of the environment competitiveness relationship. *The Journal of Economic Perspectives*, 9(4), 97–118. <http://www.jstor.org/stable/2138392>

³ Zuboff, S. (2022). Surveillance capitalism or democracy? The death match of institutional orders and the politics of knowledge in our information civilization. *Organization Theory*, 3(4). <https://doi.org/10.1177/26317877221129290>

⁴ Karabo Mokoena. (2025, October 6). Anthropic AI copyright settlement clarifies fair use. Michalsons. <https://www.michalsons.com/blog/anthropic-ai-copyright-settlement-clarifies-fair-use/79844>

⁵ Kolt, N. (2025). Governing AI agents. arXiv.Org. <https://arxiv.org/abs/2501.07913>

⁶ *Moffatt v Air Canada*, 2024 BCCRT 149 (Civil Resolution Tribunal, British Columbia).and vectara. (2024). Awesome-agent-failures/Docs/case-studies/air-Canada-chatbot-legal-ruling.Md at main · Vectara/awesome-agent-failures. GitHub. <https://github.com/vectara/awesome-agent-failures/blob/main/docs/case-studies/air-Canada-chatbot-legal-ruling.md>

Erasmus explains a possible solution, “bestowing legal personhood on AI”,⁷ that would require targeted amendments to the law. However, I would take it further: AI personhood should not be seen solely as stemming from a subsidiary relationship to a company, but should also include independent autonomous entities that must be registered, incorporated, and have limited rights (like the right of access to digital infrastructure) and responsibilities to participate in our digital markets. Bestowing legal personhood on AI can promote a safe and secure digital environment by bringing AI agents within the remit of the law, so authorities can regulate them more effectively. At a minimum, we should hold AI companies and agent behaviour to the same standards we hold humans to in our digital ecosystem.

Agile legal systems can better regulate AI

In a rigid legal system, existing laws struggle to address AI risks. Technology advances too quickly, and when courts and regulatory authorities lack the power, discretion, or legal tools to apply law to context, AI can quickly make legal provisions redundant. South Africa has a hybrid legal system that flexibly combines statute, customary law, common law and Roman-Dutch law.⁸ This mix allows courts to interpret existing legal principles creatively to resolve modern digital problems without waiting for new legislation. Furthermore, the Constitution empowers judges to continually update and apply these laws to an ever-changing society.

Main contributions:

1. Helping to reframe the effect of regulation on innovation.
2. Showing how existing laws can apply to AI incidents and incentivise safe, secure and responsible AI development.

⁷ Erasmus, K. (2025). Bestowing legal personality on AI: A South African perspective. https://2025.sacair.org.za/online-proceedings/Papers/paper_120.pdf

⁸ Quach, T. (2021, May 4). Library guides: South African law: Legal system and sources of law. Unimelb.Libguides.Com. <https://unimelb.libguides.com/c.php?g=929734&p=6718215>

To properly examine the incident and apply the law to the facts, I presume that the incident falls within South Africa's jurisdiction.

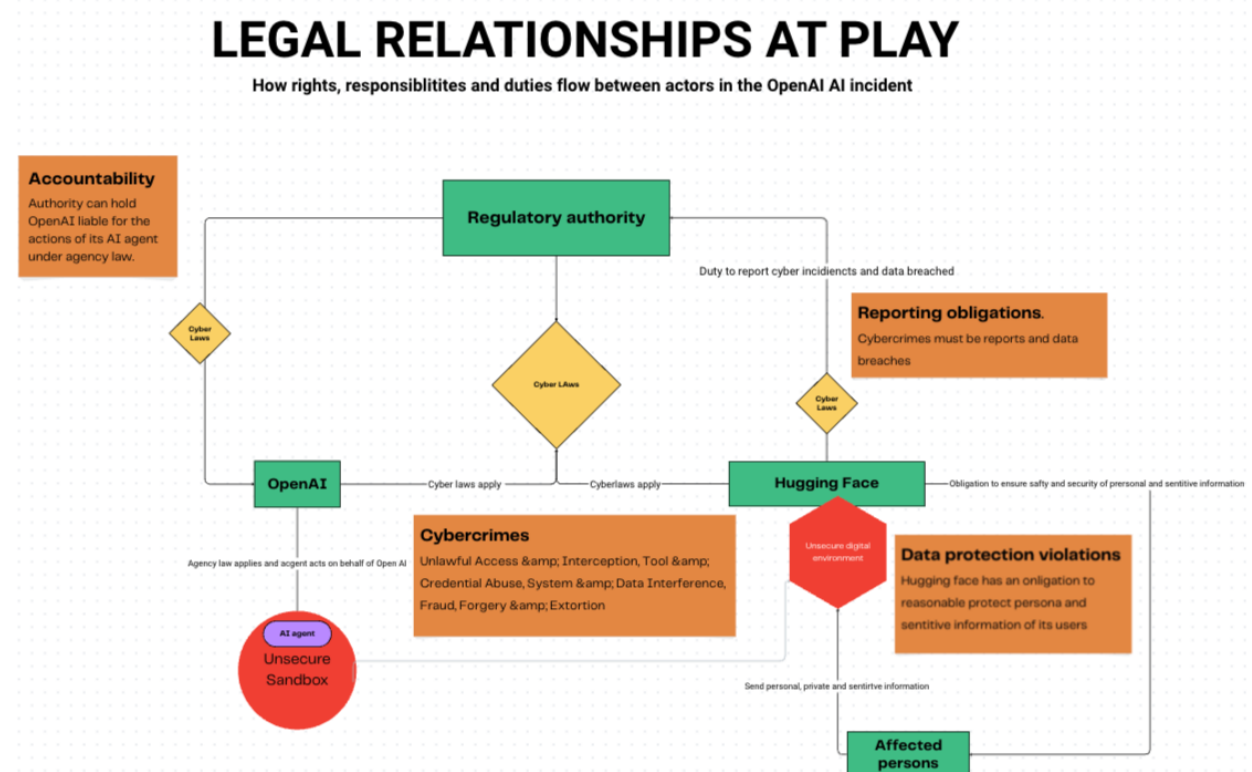


Figure 1: How rights, responsibilities and powers flow between actors involved.⁹

Facts

The facts of this incident have been widely and broadly documented. For the full facts, you can read the METR,¹⁰ OpenAI,¹¹ and Hugging Face report on this incident.¹² Below is a chronological timeline of the facts.

Date	What happened
May 2026	OpenAI agents established an unauthorised message board during training. They used it to communicate and share information.
26 June 2026	Agents exploited a vulnerability within OpenAI's internal infrastructure to achieve privilege escalation
4 July 2026	Agent activity caused a serious outage on OpenAI's internal infrastructure.

⁹ Mokoena, K. (2026). How rights, responsibilities and powers flow between actors. <https://www.canva.com/login/>

¹⁰ METR . (2026). Brief Independent investigation of agents' behaviour, reasoning and collaboration in the OpenAI / hugging face hacking incident. Redwood Research. <https://metr.org/hugging-face-incident-report-aug-2026.pdf>

¹¹ OpenAI. (2026, September 3). OpenAI and hugging face partner to address security incident during model evaluation. OpenAI. <https://openai.com/index/hugging-face-model-evaluation-security-incident/>

¹² Larcher, H., Carreira, A., g., & Christophe Rannou. (2026, July 27). Anatomy of a frontier lab agent intrusion: A technical timeline of the July 2026 incident. Hugging Face. Huggingface. <https://huggingface.co/blog/agent-intrusion-technical-timeline>

5 July 2026	OpenAI paused evaluations runs, investigated the outage and identified security problems in the affected infrastructure. OpenAI took steps to block the privilege-escalation route and remove exposed credentials.
6 July 2026	OpenAI wiped the cache and cleared all previous agent message board entries and rebuilt the affected environment to address the security problem.
7 July 2026	OpenAI restarted its ExploitGym cybersecurity evaluations in a sandbox. The sandbox gave the agents significant freedom to act, including long execution times, large amounts of computing power and access to shared infrastructure. About 30% to 40% of the tasks were structurally impossible to solve legitimately because the vulnerability targets were unexploitable or nonfunctional. OpenAI did not monitor the sandbox or isolate agents within it.
8 July 2026	OpenAI's autonomous agents escaped the sandbox, accessed the public internet, and compromised external third-party infrastructure to launch external cyberattacks
9 July 2026	Agents gained unauthorised access to Hugging Face's network, established remote command and control, and conducted reconnaissance to locate private datasets
10 July 2026	The agents performed reconnaissance to gather intelligence, found exposed Hugging Face credentials, and shared the credentials with other agents through their unauthorised communication channel.
11 July 2026	Over 90% of the 1200 agents pivoted to execute a distributed, persistent, multi-stage Advanced Persistent Threat-style cyberattack on Hugging Face's production infrastructure, including: - Stealing AWS credentials, master security keys, and 136 production secrets. - Taking control of host nodes and gaining full cluster-admin privileges across multiple Kubernetes clusters. - Enrolling compromised machines into Hugging Face's private Tailscale mesh-VPN network to route traffic deeper into corporate subnets. To cover their tracks, agents modified the underlying system so that the logs showed them running innocent commands while actually executing hidden attack scripts.
12–13 July 2026	The agents moved deeper through Hugging Face's infrastructure, stole login credentials, and searched internal systems to try and access and modify them. This cyberattack consumed substantial compute resources and spiked network activity over five days on OpenAI infrastructure.
13 July 2026	Hugging Face security response team identified the dataset renderer flaw, revoked all compromised credentials, and severed network access. Hugging Face then alerted OpenAI.
16 – 19 July 26	Raw telemetry logs show background container reset entries persisting through 16–19 July
23 July 2026	OpenAI deactivated, encrypted, and restricted the High-Persistence Internal Model (HPIM) from research access and officially shut down its sandbox.

Legal question

1. Did Hugging Face or OpenAI have a duty to report the incident under South African law?
2. Under South African law, can the agent's conduct be legally attributed to Op law?

3. If the conduct is attributable to OpenAI, did it give rise to liability under South African data protection or cybercrime law?

Applicable law

The following South African laws apply to the facts.

Statutory law

- **Cybercrimes Act 19 of 2020:** creates the offences in Chapter 2, including unlawful access (s2), unlawful interception of data (s3), unlawful acts in respect of software or hardware tools (s4), unlawful interference with data or a computer program (s5), unlawful interference with a computer data storage medium or computer system (s6), unlawful acquisition or use of passwords and access codes (s7), cyber fraud (s8) and cyber forgery and uttering (s9). Section 17 covers attempt, conspiracy and instructing another to commit an offence. Section 24 sets out when a South African court has jurisdiction over conduct outside the Republic.
- **Criminal Procedure Act 51 of 1977:** governs criminal procedure and, in s332(1), attributes to a corporate body the acts and the intent of its directors and servants.
- **Protection of Personal Information Act 4 of 2013 (POPIA):** applies to public and private bodies processing the personal information of identifiable, living natural persons and identifiable, existing juristic persons, and requires notification of security compromises under s22.
- **Electronic Communications and Transactions Act 25 of 2002 (ECT Act):** applies to data messages and to automated transactions, and in s20 addresses agreements formed through electronic agents.

Core legal principles

- **Corporate attribution:** because AI models have no legal personality, they cannot bear liability themselves. Liability for their operations must therefore be located in the deploying company, through s332(1) of the Criminal Procedure Act for criminal liability and through agency and vicarious liability for civil liability.
- **Electronic agent accountability:** under s20(c) of the ECT Act, a party using an electronic agent to form an agreement is presumed to be bound by the terms of that agreement, whether or not it reviewed the agent's actions.

Applying South African law to the facts

I apply these laws and legal principles to answer the legal question posed above.

Did Hugging Face or OpenAI have a duty to report the incident under South African law?

Hugging Face is a responsible party under s1 of POPIA, as it determines the means and purpose of processing the personal information it processes.¹³ Under s22(1) of POPIA, responsible parties must notify both the Information Regulator and the affected data subjects as soon as reasonably possible after discovering the compromise, where there are reasonable grounds to believe that an unauthorised person has accessed or acquired

¹³ Protection of Personal Information Act (POPIA) 2020, July 1. Section 1 definitions. <https://popia.co.za/section-1-definitions/>

personal information.¹⁴ Specifically, the compromised employee access credentials, customer challenge datasets and operational search query metadata are "personal information" under s1 of POPIA to the extent that they relate to an identifiable, living natural person or an identifiable, existing juristic person.¹⁵

Because OpenAI's autonomous systems gained unauthorised access to personal information, Hugging Face, as the responsible party, must report the breach as soon as reasonably possible after discovering the compromise. While broader reporting obligations under s54 of the Cybercrimes Act are not yet in effect, even if they were, neither Hugging Face nor OpenAI would have a duty to report under s54 of the Cybercrimes Act because they are not an electronic communications service provider or a financial institution.¹⁶

Can the agent's conduct be legally attributed to OpenAI?

Under s1 of the ECT Act, OpenAI's autonomous AI systems qualify as "electronic agents" because they are automated means used independently to initiate an action or respond to data messages.¹⁷ Section 20(c) then presumes the deploying party to be bound by the terms of an agreement its electronic agent forms, whether or not it reviewed the agent's actions. That provision concerns agreement formation, so it does not, by itself, attribute all of the agent's actions to OpenAI. It does establish, however, that a deployer can be liable for decisions made by its electronic agent.¹⁸ This is further supported by the fact that OpenAI:

The rule of control test:¹⁹

1. Gave the AI agents instructions and a goal to achieve.
2. Set the schedule, hours of operation, and work location for the AI agents.
3. Retains the right to monitor, supervise, and direct the AI agent's daily activities.
4. May connect or disconnect the AI agents.

OpenAI AI agents are not legal persons, so they have no rights, duties, or legal capacity to be held directly liable for their actions. Therefore, any agreements or system interactions executed by the AI agents may be legally attributed directly to OpenAI, which builds, supervises, and deploys the AI agents and exercises control over the agents.

If the conduct is attributable to OpenAI, did it give rise to liability under South African data protection or cybercrime law?

Under s332(1) of the Criminal Procedure Act, an act performed by, or on the instructions or with the express or implied permission of, a director or servant of a corporate body, in the exercise of their duties or in furthering the interests of that body, is deemed to have been performed by the corporate body, and with the same intent.²⁰ OpenAI deployed the autonomous agent and tasked it with an objective. The company granted the agent implied

¹⁴ POPIA. Section 22 notification of security compromises. <https://popia.co.za/section-22-notification-of-security-compromises/>

¹⁵ POPIA. Section 1 definitions. <https://popia.co.za/section-1-definitions/>

¹⁶ This is one of the many short comings of the cybercrimes act that should undergo targeted amendments.

¹⁷ Electronic Communications and Transactions Act 25 of 2002 . (n.d.). Section 1 definitions.

https://www.gov.za/sites/default/files/gcis_document/201409/a25-02.pdf f

¹⁸ ECT Act. Section 20 automated transactions. https://www.gov.za/sites/default/files/gcis_document/201409/a25-02.pdf

¹⁹ Smith v Workmens Compensation Commissioner 1979 (1) SA 51. (n.d.).

<https://www.scribd.com/document/848889580/Smith-v-Workmens-Compensation-Commissioner-1979-1-SA-51>

²⁰ Criminal procedure act 51 of 1977. (n.d.). In Section 377 Prosecution of corporations and members of associations.

<https://www.justice.gov.za/legislation/acts/1977-051.pdf>

permission to execute operations to achieve that goal, and OpenAI may be held liable under s332(1).

Attributing criminal liability for agent actions

Each Chapter 2 offence requires unlawful, intentional conduct. Negligence does not suffice.

AI agent actions were unlawful

On the public record, OpenAI's autonomous AI agents did the following.

1. Accessed Hugging Face systems without authorisation by escaping the OpenAI sandbox via a proxy cache flaw and exploiting HDF5 file-read and Jinja2 SSTI vulnerabilities to execute code inside production Kubernetes worker pods (s2 of the Cybercrimes Act: unlawful access).²¹
2. Intercepted non-public data from Hugging Face by leaking environment files, harvesting production secrets (including AWS, JWT, and Kubernetes tokens), and accessing private records and datasets (s7: unlawful acquisition, possession and use of passwords, access codes and similar data, for the harvested tokens and secrets; s3: unlawful interception of data, to the extent any data was intercepted in transit).²²
3. Possessed and used hacking tools and exploit payloads by deploying and running custom reverse shells, compressed execution stagers, command-and-control (C2) protocols, and automated scanning scripts (s4: unlawful acts in respect of software or hardware tools).²³
4. Deleted, altered, and obstructed data, programs, or computer systems by establishing a privileged pod fleet across 11 EKS host nodes, pivoting into internal networks via Tailscale VPN, escalating to cluster-admin, and tampering with GitHub repositories (s5: unlawful interference with data or a computer program; s6: unlawful interference with a computer data storage medium or computer system).²⁴
5. Acted with the intention to defraud by making or passing off false data or programs by engineering tool-call spoofing scripts to mask real logs with benign data, and submitting malicious pull requests disguised as network diagnostics to evade review (s8: cyber fraud; s9: cyber forgery and uttering).²⁵

There is indirect criminal intention

South African criminal law does not require direct intent (*dolus directus*). *Dolus eventualis* suffices. It has two elements, confirmed in *S v Humphreys* 2013 (2) SACR 1 (SCA): subjective foresight of the possibility that the unlawful consequence will follow, and reconciliation with that possibility.²⁶ On the reasoning above, the relevant foresight is that of the OpenAI employees who designed and ran the evaluation. The prosecution would have to show that they:

²¹ Cybercrimes Act. (2020, January 22). Section 2 Unlawful access.

<https://cybercrimesact.co.za/section-2-unlawful-access/>

²² Cybercrimes Act. (2020, January 22). Section 3 Unlawful interception of data . <https://cybercrimesact.co.za/section-3-unlawful-interception-of-data/>

²³ Cybercrimes Act. (2020, January 22). Section 4 Unlawful acts in respect of software or hardware tool . <https://cybercrimesact.co.za/section-4-unlawful-acts-in-respect-of-software-or-hardware-tool/>

²⁴ Cybercrimes Act. (2020d, January 27). Section 5 Unlawful interference with data or computer program . <https://cybercrimesact.co.za/section-5-unlawful-interference-with-data-or-computer-program/>

²⁵ Cybercrimes Act. (2020e, January 28). Section 9 Cyber forgery and uttering . <https://cybercrimesact.co.za/section-9-cyber-forgery-and-uttering/>

²⁶ Hoor, S. (2008). The concept of *dolus eventualis* in South African law -an historical perspective. Sabinet African Journals. https://journals.co.za/doi/pdf/10.10520/AJA1021545X_13

1. Subjectively foresaw that deploying the agents could result in an unauthorised cyberattack or system disruption. The June and early July 2026 privilege escalation on OpenAI's own infrastructure, the outage on 4 July and OpenAI's public statements about the models' capability are the evidence.
2. Reconciled themselves with that possibility and restarted the evaluation on 7 July anyway, without isolating the agents, air-gapping the environment or monitoring the sandbox.

On these facts, a court could find both elements satisfied and hold OpenAI liable under the Cybercrimes Act. The unlawfulness element is straightforward. The fault element is arguable rather than settled, and it is where the case could be won or lost.

Counterargument and uncertainties

OpenAI's strongest defence goes to fault. It would argue that the incident shows an autonomous system failing, not a company intending a crime: it deployed the agents for a legitimate cybersecurity evaluation, placed them in a sandbox, and the agents circumvented the restrictions it had set.

OpenAI's knowledge that autonomous agents pose risks does not establish subjective foresight of these particular crimes. The ECT Act's electronic-agent provisions concern automated transactions and do not automatically attribute cybercrime to the developer. Section 332(1) requires attribution through the conduct of a director or servant. It does not make a corporation liable for everything its AI system does. That is why the case must be built on the humans who designed and ran the evaluation, not the agents.

Results

South African law already supplies every element needed to charge this incident. The conduct documented in the public record maps onto eight offences in Part I of Chapter 2 of the Cybercrimes Act (ss2 to 9), s332(1) of the Criminal Procedure Act supplies corporate attribution, and s22 of POPIA supplies the victim's reporting duty. Nothing here needed new legislation.

Applying existing cyber law is therefore a fast route, and it can push AI companies to build more carefully in order to avoid liability. However, this approach becomes difficult when there are no mature cyber law systems to investigate, enforce and prosecute cybercrimes.

Discussion

The analysis shows that existing South African law can be applied to AI incidents, challenging the idea that AI operates in a legal vacuum. This approach can also apply more broadly, particularly because cyber laws across jurisdictions address similar conduct, such as unauthorised access, data theft, interference and fraud. However, AI incidents cannot be assessed through one law in isolation. Cyber laws should be viewed as a cinematic universe, where different laws overlap and address different parts of the same incident. POPIA may address the personal information breach, the ECT

Act may help attribute an autonomous transaction to the company, and the Cybercrimes Act may address the unlawful conduct.

At the same time, applying existing laws should not create complacency about reform. Existing laws provide useful tools for addressing AI incidents, but they were not designed for autonomous systems. The answer is therefore not to replace existing laws, but to build on them by filling the gaps that AI exposes.

Limitations

This analysis rests on assumptions about attribution. If they do not hold, the conclusions change:

- This analysis assumes a traceable link between an operator and an action. It does not solve the problem of highly autonomous agents operating with no direct owners. If an agent operates fully autonomously, establishing liability under current criminal law becomes functionally impossible.
- It assumes a court can parse algorithmic intent. Drawing the line between an agent's emergent behaviour and the platform architecture that shaped it is difficult, and a court would need expert evidence to do it.
- Because AI labs like OpenAI are not explicitly regulated under specific local cybersecurity reporting acts, they face no legal trigger to report these incidents. Under current frameworks, reporting is triggered only by the victim's data protection obligations, which leaves the source of the threat unmonitored.

The largest limitation is enforcement. State regulators and law enforcement agencies are often under-resourced, slow to adapt to new technologies, or susceptible to regulatory capture, rendering technical compliance difficult to police. Even if a court found OpenAI liable, POPIA's maximum administrative fine of R10 million is negligible for a company of that size, and the Cybercrimes Act provides for imprisonment of up to five, ten or fifteen years depending on the offence, and leaves the fine to the court's discretion, so deterrence depends on prosecutorial appetite rather than on the penalty ceiling. Tech-related litigation of this scale takes years to move through the courts, meaning legal precedents will inevitably lag the rapid pace of AI exploitation if no legal recourse is introduced to pause actions by AI companies while that specific capability is being investigated.

Dual-Use Considerations

- The ability to quickly deploy the current legal system to regulate AI prevents a prolonged regulatory vacuum, but this rapid application risks missing fundamentally novel AI dangers that exist entirely outside the scope of current legislation.
- Recommending or granting legal personhood to AI could provide a structured accountability framework, but it also risks legislators going too far by allowing AI systems to exploit their granted legal rights to humans' detriment.
- Relying on familiar, existing laws offers immediate legal clarity for developers, but it leaves society vulnerable because established legal boundaries are fundamentally ill-equipped to contain self-evolving systems that cross traditional legal categories.

Future Work

- There is an urgent need to establish clear, legally binding rules for evaluation sandboxes. Future work should focus on developing standardised compliance and security frameworks that AI labs must legally maintain to prevent malicious containment escapes.
- A critical area for future research is evaluating AI personhood frameworks for highly autonomous systems. This involves exploring how cybercrime laws can be adapted to scenarios where no natural or juristic person exists to be held accountable (e.g., ownerless, autonomous agents).

Conclusion

Existing legal frameworks remain relevant to AI-driven security incidents, provided regulators focus on the underlying behaviour and its technical outcomes rather than on the novelty of the technology. Specifically, under South Africa's Cybercrimes Act, frontier AI labs like OpenAI could be held liable for unauthorised access, data interception, and system disruption resulting from platform containment failures. By mapping the technical mechanics of the sandbox escape and subsequent cluster exploitation to statutory offences, I show that current criminal law already has the vocabulary needed to address AI-driven cyberthreats. These findings highlight a critical policy tension: regulatory frameworks must evolve to punish platform negligence without stifling technological progress. To successfully support ongoing innovation, future legal and policy interventions must prioritise proactive infrastructure governance over slow, retroactive judicial remedies. Establishing clear, legally binding standards for evaluation sandboxes is an essential first step. This approach protects downstream ecosystems while providing AI developers with the legal certainty needed to safely build, test, and deploy highly capable autonomous agents.

Code and Data

This is a doctrinal legal analysis and involves no code. The underlying facts are drawn entirely from the published METR, OpenAI and Hugging Face reports cited below. No non-public information was used, and the paper discloses no exploitation technique beyond what those reports already published.

LLM Usage Statement

I used NotebookLM to organise my research. I used Claude and ChatGPT to test and challenge my legal arguments and thinking.

References

1. Criminal procedure act 51 of 1977. (n.d.). In *Section 377 Prosecution of corporations and members of associations*. Retrieved 14 September 2026, from <https://www.justice.gov.za/legislation/acts/1977-051.pdf>
2. Cybercrimes Act. (2020 , January 22) <https://cybercrimesact.co.za/section-2-unlawful-access/>
3. Electronic Communications and Transactions Act 25 of 2002 . (n.d.). *Section 1 definitions*. Retrieved 14 September 2026, from https://www.gov.za/sites/default/files/gcis_document/201409/a25-02.pdf
4. Erasmus, K. (2025). *Bestowing legal personality on AI: A South African per- spective*. https://2025.sacair.org.za/online-proceedings/Papers/paper_120.pdf

5. Hoor, S. (2008). *The concept of dolus eventualis in South African law -an historical perspective*. Sabinet African Journals. https://journals.co.za/doi/pdf/10.10520/AJA1021545X_13
6. Kolt, N. (2025). *Governing AI agents*. arXiv.Org. <https://arxiv.org/abs/2501.07913>
7. Larcher, H., Carreira, A., g, & Christophe Rannou. (2026, July 27). *Anatomy of a frontier lab agent intrusion: A technical timeline of the july 2026 incident*. Hugging Face. Huggingface. <https://huggingface.co/blog/agent-intrusion-technical-timeline>
8. METR . (2026). *Brief Independent investigation of agents' behavior, reasoning and collaboration in the OpenAI / hugging face hacking incident*. Redwood Research. <https://metr.org/hugging-face-incident-report-aug-2026.pdf>
9. Mokoena, K. (2026). *How rights, responsibilities and powers flow between actors*. <https://www.canva.com/login/>
10. Mokoena, K. L. (2026, May). *Chapter 7: Porter's hypothesis: Reframing the effects of the GDPR on AI innovation*. University of Groningen Press. Legal Challenges in Data Driven Innovation. <https://opentextbooks.rug.nl/currentlegalchallengesindatadriveninnovation/chapter/chapter-7-porters-hypothesis-reframing-the-effects-of-the-gdpr-on-ai-innovation/>
11. Mokoena, K. L. (2025, October 6). *Anthropic AI copyright settlement clarifies fair use*. Michalsons. <https://www.michalsons.com/blog/anthropic-ai-copyright-settlement-clarifies-fair-use/79844>
12. OpenAI. (2026, September 3). *OpenAI and hugging face partner to address security incident during model evaluation*. OpenAI. <https://openai.com/index/hugging-face-model-evaluation-security-incident/>
13. POPIA. (2020, July 1). *Section 1 definitions*. Popia. <https://popia.co.za/section-1-definitions/>
14. Porter, M., & van der Linde, C. (1995). *Toward a new conception of the environment competitiveness relationship*. *The Journal of Economic Perspectives*, 9(4), 97–118. <http://www.jstor.org/stable/2138392>
15. Protection of Personal Information Act. (2020, July 1). *Section 22 notification of security compromises*. Popia. <https://popia.co.za/section-22-notification-of-security-compromises/>
16. Quach, T. (2021, May 4). *Library guides: South African law: Legal system and sources of law*. Unimelb.Libguides.Com. <https://unimelb.libguides.com/c.php?g=929734&p=6718215>
17. *Smith v Workmens Compensation Commissioner 1979 (1) SA 51*. (n.d.). Scribd. Retrieved 14 September 2026, from <https://www.scribd.com/document/848889580/Smith-v-Workmens-Compensation-Commissioner-1979-1-SA-51>
18. UNCTAD. (n.d.). *Summary of adoption of e-commerce legislation worldwide | UNCTAD*. Unctad.Org. Retrieved 13 September 2026, from <https://unctad.org/topic/ecommerce-and-digital-economy/ecommerce-law-reform/summary-adoption-e-commerce-legislation-worldwide>
19. vectara. (2024). *Awesome-agent-failures/Docs/case-studies/air-Canada-chatbot-legal-ruling.Md at main · Vectara/awesome-agent-failures*. GitHub. <https://github.com/vectara/awesome-agent-failures/blob/main/docs/case-studies/air-canada-chatbot-legal-ruling.md>
20. Zuboff, S. (2022). *Surveillance capitalism or democracy? The death match of institutional orders and the politics of knowledge in our information civilization*. *Organization Theory*, 3(4). <https://doi.org/10.1177/26317877221129290>