Written Evidence submitted by PauseAI UK (CSRB09)

Cyber Security and Resilience Bill

Prepared by PauseAI UK – January 2026

About PauseAI UK

1. PauseAI UK is an organisation advocating for safer artificial intelligence (AI) development, part of the PauseAI Global network. We assisted in the organisation of the December 2025 Westminster Hall debate on AI safety[1] and coordinated a cross-party letter on the Frontier AI Safety Commitments signed by over 60 British politicians.[2]

2. Our team of organisers and volunteers includes AI safety researchers, software engineers, and technology lawyers. Our advocacy for pausing frontier AI development stems from the same evidence base we draw on here: the documented risks of autonomous and adaptive systems operating in high-stakes environments. We support measures that reduce these risks, whether through development control or improved resilience of deployed systems.

3. We are motivated by the severe risks of future AI systems. Our contribution to this Committee is the result of sustained attention to how autonomous and adaptive machine learning systems actually behave, drawing on government, academic, and industry sources. The failure modes we describe are unintuitive, and therefore can be easily missed, because the systems themselves operate in surprising ways.

4. We welcome this opportunity to provide written evidence to the Committee and hope it will prove useful to its deliberations. We remain at the Committee's disposal.

Summary

5. This evidence addresses gaps in the Cyber Security and Resilience Bill's treatment of risks arising from AI, specifically autonomous and adaptive machine learning (ML) systems[3]. These systems are increasingly integrated into critical national infrastructure but present failure modes that existing incident management frameworks do not adequately capture.

6. The Bill's stated purpose of improving Critical National Infrastructure (CNI) resilience is sound and will be increasingly important in a world with sophisticated AI systems. The amendments we propose address remaining AI-related vulnerabilities which we believe will be

[1] Hansard, HC Deb, 10 December 2025, AI Safety debate. https://hansard.parliament.uk/commons/2025-12-10/debates/9F01B4B9-12CB-42E2-84E2-A65F7D30BFAF/AISafety

[2] PauseAI UK, "Dear Sir Demis" letter, December 2025. https://pauseai.info/dear-sir-demis-2025

[3] We largely refer to "autonomous or adaptive ML systems" rather than "artificial intelligence." This terminology is more technically precise and aligns with the properties that create the distinct risks we describe, while excluding many ‘AI’ systems which pose no additional threat. We are concerned about systems whose behaviour emerges from training rather than explicit specification, and which can operate or adapt without ongoing human guidance. In our proposed amendment text, we use the fuller phrase "autonomous or adaptive systems based on machine learning" for legislative clarity.

difficult to anticipate without sustained attention on the use and behavior of autonomous and adaptive ML systems.

Key concerns:

7. (a) **Incident reporting provisions do not capture novel failure modes** from autonomous or adaptive ML systems. Current criteria for reportable incidents focus on "significant impact" in a purely quantitative manner (e.g. tracking geographical scope, number of effected users etc). These criteria miss an important qualitative risk factor: autonomous or adaptive ML systems go wrong in unexpected and unpredictable ways, and this unpredictability is a hallmark of significant risk in itself.

8. (b) **No requirement to share incident intelligence with the UK AI Security Institute** (AISI), the UK's expert body on advanced AI security. AISI cannot build threat pattern recognition without systematic incident data.

9. (c) **Secretary of State reporting occurs at intervals of up to five years**, which is insufficient given the pace of adoption of ML systems in global infrastructure and of capability change; the AISI *Frontier AI Trends Report* documents relevant capabilities doubling roughly every eight months.[4]

10. (d) **No requirement for a code of practice addressing autonomous or adaptive ML systems** in critical infrastructure, despite DSIT publishing voluntary guidance in this area.

11. (e) **Inadequate framework for safe shutdown of compromised AI systems.** Two gaps: (i) nothing requires the technical capability to disable autonomous or adaptive ML systems to exist before a crisis; (ii) no requirement for fallback capability, creating perverse incentives to keep compromised systems running rather than face service blackout and leaving infrastructure exposed if adversarial actors or the systems themselves trigger disruption. We consider this the most important vulnerability for the Committee to examine.

12. (f) **Critical suppliers face designation without corresponding baseline duties**, leaving critical supply chain concentration risks unaddressed by default.

13. (g) **Public sector exclusion creates an accountability gap.** The Bill's regulatory scope appropriately focuses on private sector CNI, but the existing Section 40 reporting requirement could be amended to create parliamentary visibility of public sector cyber resilience without regulatory expansion.

14. **We consider the incident intelligence provisions (concerns a-b) and operational continuity requirement (concern e) most urgent.**

15. We propose amendments to Sections 15, 18, 36, 40, 12, and 29, as detailed below.

[4] AI Security Institute, Frontier AI Trends Report, December 2025, p. 10, Figure 3. https://www.aisi.gov.uk/frontier-ai-trends-report

1. Incident intelligence for autonomous and adaptive ML systems

16. **VULNERABILITY:** Autonomous and adaptive ML systems operate by carrying out their own analyses based on training processes; designers cannot robustly specify outcomes at design time. Incident reporting criteria capture impact metrics but not ML-specific failure patterns, preventing effective pattern recognition. AISI, the UK's expert body on advanced AI security, receives no incident data under the Bill's framework.

17. **THREAT MODEL:** An autonomous or adaptive ML system deployed in CNI fails in an unexpected way. Under current reporting criteria, the incident is classified by its effects (disruption extent, users affected, duration) rather than its cause. The novel failure mode goes unrecognised as a pattern. The same model architecture is deployed elsewhere. The vulnerability reappears, with consequences proportional to the function and degree of autonomy and adaptiveness of the deployed system.

Current framework:

18. Incident occurs ⇒ Criteria applied: disruption extent, users affected, duration, geographic scope ⇒ [gap: no ML-specific flag] ⇒ Novel failure mode classified as ordinary incident ⇒ No technical pattern recognition ⇒ Same failure mode re-occurs despite now being predictable.

With proposed amendment:

19. Incident occurs ⇒ Criteria applied: existing criteria + ML system involvement ⇒ Regulator receives flagged report ⇒ CSIRT notifies AISI of ML-involved incidents ⇒ AISI analyses for patterns ⇒ Prompt warning across CNI sectors ⇒ Higher likelihood of preventing similar harm.

Background

20. During the Commons second reading, Oliver Dowden MP warned about "the risk of cyber-attacks posed by agentic artificial intelligence," noting the UK may be "in the foothills of the risk posed by agentic AI."[5] The NCSC has assessed that AI will "almost certainly" increase both the volume and impact of cyber attacks.[6]

21. MI5 Director General Sir Ken McCallum stated: "[w]e also need to scope out the next frontier: potential future risks from non-human, autonomous AI systems which may evade human oversight and control ... It would be reckless to ignore the potential for AI to cause harm."[7]

22. Current incident significance criteria (Regulations 11, 11A, 12A, 14E) assess impact through factors including disruption extent, user numbers, duration, and geographic area,

[5] https://hansard.parliament.uk/commons/2026-01-06/debates/BB815F91-651E-4A24-AAFE-8BD7D92B2033/CyberSecurityAndResilience%28NetworkAndInformationSystems%29Bill#contribution-E730FA1F-E3A8-4A79-B28E-4FFB83D9A5E3

[6] NCSC, Impact of AI on Cyber Threat from Now to 2027, May 2025, p. 2. https://www.ncsc.gov.uk/pdfs/report/impact-ai-cyber-threat-now-2027.pdf

[7] MI5, "Director General Ken McCallum gives threat update," October 2025. https://www.mi5.gov.uk/director-general-ken-mccallum-gives-threat-update

understandably so, since these criteria have been fit for purpose for the forms of technology used up until recently in CNI.

23. The Government Cyber Action Plan adopts "scal[ing] what works"[8] as its core approach. What has worked for deterministic systems may not scale to unpredictable and potentially agentic technology. The criteria that served CNI when systems behaved predictably need thoughtful augmentation for systems whose behaviour emerges from training.

24. Traditional software fails through bugs in specified code. Engineers can trace cause through explicit logic, from input to processing steps to output. The behaviour of ML systems emerges from training data and learned parameters rather than explicit specification. When such systems fail, the cause may be a statistical pattern which no human specified or reviewed. Standard debugging does not apply to learned representations.

25. This assessment is shared across UK Government research bodies. David Dalrymple, Programme Director for ARIA's £59 million Safeguarded AI programme, stated: "AI could unlock transformative improvements in our critical infrastructure, but ... without ironclad safety assurances, we risk unintended and damaging consequences."[9] ARIA's programme specifically addresses the challenge of deploying AI systems in safety-critical infrastructure, the same challenge this Bill must address for CNI.

26. The AISI *Frontier AI Trends Report* provides evidence on the trajectory of these systems:

- Models now complete expert-level cyber tasks (typically requiring 10+ years of human experience).[10]

- AISI reports having found "universal jailbreaks for every system we've tested."[11] The efficacy of safeguards varies substantially between models and misuse categories.

- Autonomous task completion has expanded dramatically: in late 2023, models could almost never complete software tasks that would take a human expert over an hour (under 5% success); by mid-2025, they succeeded over 40% of the time.[12]

27. The *International AI Safety Report*, commissioned by the UK Government and chaired by Yoshua Bengio with contributions from 96 AI experts across 30 countries, identified the threat of "capabilities for autonomously using computers, programming, gaining unauthorised access to digital systems, and identifying ways to evade human oversight."[13] The October 2025 Key Update notes that "some AI systems have demonstrated strategic behaviour while being evaluated, raising potential oversight challenges."[14]

[8] Government Cyber Action Plan, January 2025. https://www.gov.uk/government/publications/government-cyber-action-plan

[9] ARIA, Safeguarded AI Programme. https://www.aria.org.uk/safeguarded-ai/

[10] AI Security Institute, Frontier AI Trends Report, December 2025, p. 3, Figure 1.2. https://www.aisi.gov.uk/frontier-ai-trends-report

[11] AI Security Institute, Frontier AI Trends Report, December 2025, p. 24. https://www.aisi.gov.uk/frontier-ai-trends-report

[12] AI Security Institute, Frontier AI Trends Report, December 2025, p. 9, Figure 2. https://www.aisi.gov.uk/frontier-ai-trends-report

[13] Yoshua Bengio et al, International AI Safety Report, January 2025, p. 19. https://internationalaisafetyreport.org

[14] Yoshua Bengio et al, International AI Safety Report 2025: First Key Update, October 2025, p. 5. https://internationalaisafetyreport.org/

28. As early as August 2023, the Centre for Emerging Technology and Security (CETaS) and Centre for Long-Term Resilience (CLTR), in their joint briefing paper *Strengthening Resilience to AI Risk: A guide for UK policymakers*, concluded that "the UK is inadequately resilient to the risks posed by AI" and called for "a model reporting and information sharing regime between AI developers and regulators."[15]

29. The Bill creates information-sharing gateways between regulated entities, competent authorities, and NCSC. In the Bill's second reading, Minister Ian Murray expressed the aim, among other objectives, to "not only give us better information, but help agencies to warn others, should they need to, before they become the next targets."[16] It does not create any pathway for incident intelligence to reach AISI. AISI's mission includes understanding how advanced AI systems behave in deployment. CNI incident data would directly support this mission. Under the current framework, AISI must rely on voluntary disclosure or public reporting to understand how AI systems are failing in critical infrastructure, precisely the context where such failures matter most. Our own work highlighted last year that relying on voluntary information sharing on AI systems was insufficient for AISI to obtain access to the relevant information.[17]

Recommendations

30. We recommend adding autonomous or adaptive ML system involvement to the incident significance criteria. This enables collection of data on failure patterns before catastrophic incidents occur, directly supporting the infrastructural resilience the Bill seeks to create.

31. We also recommend requiring the CSIRT to share relevant incident information with AISI, creating a systematic intelligence pipeline for novel threats. This should create minimal additional burden (a single additional reporting element for CSIRT when processing relevant incidents) while enabling AISI to build the cross-sectoral intelligence necessary to anticipate emerging threats.

OVERVIEW OF PROPOSED AMENDMENT - Note: the full proposed amendment text for this recommendation is available as an appendix (pp. 20-23).

– Part 1: Section 15 (incident significance criteria)

32. Add to the factors determining whether an incident is "significant":

> *"(x) whether the incident involves failure modes not previously observed in the relevant sector materially involving autonomous or adaptive systems based on machine learning, including where the potential impact of such failure modes was mitigated or prevented."*

[15] Ardi Janjeva et al, Strengthening Resilience to AI Risk: A guide for UK policymakers, CETaS Briefing Papers, August 2023. https://cetas.turing.ac.uk/publications/strengthening-resilience-ai-risk

[16] https://hansard.parliament.uk/commons/2026-01-06/debates/BB815F91-651E-4A24-AAFE-8BD7D92B2033/CyberSecurityAndResilience(NetworkAndInformationSystems)Bill#contribution-E997FE54-0549-4710-B15A-3A055AE7F1B8

[17] Following a letter issued by PauseAI UK, TIME confirmed that DeepMind did not give pre-release access to Gemini 2.5 Pro, August 2025. https://time.com/7313320/google-deepmind-gemini-ai-safety-pledge/

33. Add to reportable incident information:

> *"(x) where the incident was associated with one or more autonomous or adaptive systems based on machine learning, details of those systems and their involvement in the incident."*

34. Note on interpretation: "failure modes not previously observed" is intended to capture incidents that appear novel to the reporting entity based on available information. We anticipate regulators would maintain and share a registry of known ML failure modes, enabling entities to make good-faith assessments. The criterion should err toward inclusion: where doubt exists about whether a failure mode has been previously observed, the incident should be reported.

– Part 2: Section 18 (Information sharing)

35. Insert new Regulation 6C:

> *"**Sharing of information with the AI Security Institute***
>
> *6C.—(1) Where the CSIRT receives notification of an incident under regulation 11, 11A, 12A, or 14E that materially involves autonomous or adaptive systems based on machine learning, the CSIRT must share relevant technical information with the AI Security Institute within 72 hours.*
>
> *(2) In this regulation, 'AI Security Institute' means such body as the Secretary of State may designate for the purposes of this regulation."*

36. Note on interpretation: "materially involves" is intended to capture incidents where the autonomous or adaptive system's behaviour was a proximate cause or significant contributing factor, as distinct from incidents where such systems were merely present. We anticipate this would be clarified in guidance issued under the Act.

2. Insufficient Secretary of State reporting frequency

37. **VULNERABILITY:** the five-year reporting cycle is mismatched with the pace of capability change in autonomous and adaptive ML systems. For example, AISI evaluations show self-replication success rates rising from under 5% in 2023 to over 60% in 2025[18] - the ability to propagate and persist without human authorisation, a capability that compounds other security risks. Vulnerabilities that did not exist when the Act was passed may become significant well before the end of a single reporting cycle, leaving Parliament without timely information to assess whether the Act's provisions remain adequate, thus exposing the UK to failure modes for which legislation is ill equipped.

Background

38. Section 40 requires the Secretary of State to report on the Act's implementation at intervals no greater than five years. The AISI *Frontier AI Trends Report* documents that "the duration of cyber tasks that AI systems can complete without human direction is roughly doubling every eight months"[19], suggesting a fast pace of development should be expected.

39. Even if this pace of capability growth moderates, the underlying trajectory is clear: the NCSC reports a 130% year-on-year increase in nationally significant incidents, with "highly significant" incidents rising 50% for the third consecutive year.[20] The *International AI Safety Report* notes: "the field of AI is moving too quickly for a single yearly publication to keep pace. Significant changes can occur on a timescale of months, sometimes weeks."[21] The NCSC concurs: “there will almost certainly be a digital divide between systems keeping pace with AI-enabled threats and a large proportion that are more vulnerable,” such that “keeping pace with 'frontier AI' capabilities will almost certainly be critical to cyber resilience for the decade to come.”[22] Accordingly, annual or biennial reporting would be more appropriate than quinquennial regardless of the precise pace of development.

Recommendations

40. We recommend reducing the maximum reporting interval to two years and, for increased flexibility and efficient horizon scanning, requiring an interim review mechanism when the Secretary of State or AISI identifies significant AI capability advances.

PROPOSED AMENDMENT – Section 40

41. Replace "5 years" with "2 years" in Section 40(1), and add:

[18] AI Security Institute, Frontier AI Trends Report, December 2025, p. 30, Figure 16. https://www.aisi.gov.uk/frontier-ai-trends-report

[19] AI Security Institute, Frontier AI Trends Report, December 2025, p. 6, Figure 3. https://www.aisi.gov.uk/frontier-ai-trends-report

[20] NCSC Annual Review 2025. https://www.ncsc.gov.uk/annual-review-2025

[21] Yoshua Bengio et al, International AI Safety Report 2025: First Key Update, October 2025, p. 4. https://internationalaisafetyreport.org/

[22] NCSC, Impact of AI on Cyber Threat from Now to 2027, May 2025, p. 2. https://www.ncsc.gov.uk/pdfs/report/impact-ai-cyber-threat-now-2027.pdf

"(1A) The Secretary of State must, within twelve months of this section coming into force, publish criteria for determining whether changes in the capabilities of autonomous or adaptive systems based on machine learning warrant an interim report under this section.

(1B) The Secretary of State must publish an interim report under this section within twelve months of—

(a) publication by the AI Security Institute, or such body as the Secretary of State may designate, of an assessment that the criteria published under subsection (1A) have been met, or

(b) the Secretary of State determining that those criteria have been met.

(1C) Each report under this section must include a review of whether the criteria published under subsection (1A) remain appropriate, and the Secretary of State must publish any revisions to those criteria."

3. No requirement for a code of practice addressing autonomous or adaptive ML systems

42. **VULNERABILITY:** voluntary guidance exists. However, there is no pathway to enforceable standards for autonomous or adaptive ML systems in CNI.

43. **THREAT MODEL:** operators deploy ML systems in CNI following general cybersecurity guidance. General guidance does not address ML-specific failure modes (e.g. learned behaviour drift, challenges related to goal specification, adversarial manipulation, emergent capabilities). Concrete vulnerabilities persist which a ML-specific code may have addressed. This undermines confidence and stability.

Background

44. DSIT has published a voluntary AI Cybersecurity Code of Practice[23] and Minister Feryal Clark stated the UK is "setting global benchmarks for secure innovation,"[24] but the Bill does not create any pathway from voluntary guidance to enforceable standards for AI systems in critical infrastructure.

45. NCSC Chief Technology Officer Ollie Whitehouse noted: "it is vital that we harness the transformative potential of AI securely so that our society can reap the benefits of new technologies without introducing avoidable vulnerabilities and cyber risks."[25]

46. Section 38 establishes that compliance with a code of practice issued under this Act is admissible as evidence of compliance with corresponding duties. This creates a meaningful incentive to follow the code. Voluntary codes do not create this evidential presumption.

47. The *International AI Safety Report* identifies concentration risk as a structural vulnerability: where multiple critical sectors depend on a small number of AI system providers, a single vulnerability can propagate failures across all dependent infrastructure simultaneously.[26] This concern applies whether the dependency is on general-purpose foundation models or on sector-specific systems where few providers dominate particular CNI applications.

48. Cross-sectoral concentration means a cross-sectoral code of practice is appropriate. ML failure modes (e.g. opacity, emergent behaviour, distribution shift) are similar regardless of whether the system is deployed in energy, healthcare, or transport. A sector-by-sector approach would create fragmentation where consistency is needed.

[23] DSIT, AI Cyber Security Code of Practice, January 2025. https://www.gov.uk/government/publications/ai-cyber-security-code-of-practice

[24] Gov.uk, "World-leading AI cyber security standard to protect digital economy," January 2025. https://www.gov.uk/government/news/world-leading-ai-cyber-security-standard-to-protect-digital-economy-and-deliver-plan-for-change

[25] Ibid.

[26] Yoshua Bengio et al, International AI Safety Report, January 2025, p. 123. https://internationalaisafetyreport.org/

Recommendations

49. We recommend requiring the Secretary of State to consider issuing a code of practice specifically addressing network and information systems that incorporate autonomous or adaptive ML capabilities. This completes work the Government has already started by issuing the AI Cybersecurity Code of Practice.

PROPOSED AMENDMENT – Section 36

50. Add new subsection:

> *"(X) The Secretary of State must, within 12 months of this section coming into force, consider whether to issue a code of practice under this section addressing the security of network and information systems that incorporate autonomous or adaptive systems based on machine learning, and must publish the reasons for any decision not to issue such a code."*

51. Note*:* we propose 12 months as a timeframe that we expect to be feasible without the issuance being too slow to be effective. The Government has already developed voluntary guidance; the question is merely whether to give it statutory footing. This should enable a timely process.

4. Inadequate framework for safe shutdown of compromised systems

52. **VULNERABILITY:** the Bill lacks an adequate framework for safe shutdown of compromised systems. Two gaps exist: (a) nothing requires the technical capability to disable autonomous or adaptive ML systems to exist before a crisis, and as such systems become more agentic, shutdown may become technically harder; and (b) the Bill currently contains no requirement for fallback capability to maintain essential functions during and after disabling. Without addressing these gaps, the safe shutdown of compromised autonomous or adaptive ML systems may be technically impossible, operationally catastrophic, or both.

53. **THREAT MODEL:** Two gaps determine whether any shutdown is survivable.

Technical shutdown capability (Gap A): an operator identifies anomalous behaviour in an autonomous or adaptive ML system and wants to disable it. Nothing in the Bill requires the technical capability to do so to exist. As such systems become more agentic, with capabilities for self-replication and distributed execution, technical shutdown may become harder.

Fallback capability (Gap B): the operator has the technical capability to disable the system, but no fallback (degraded service provision mode) exists. Disabling causes service blackout. The operator is incentivised to keep the compromised system running and hope for a mild outcome given that the alternative is visible, immediate harm.

These gaps matter regardless of why shutdown occurs: operator decision for safety reasons, authority direction under Section 43, system failure, or circumstances beyond operator control such as supply chain disruption. Addressing these gaps can provide defensive depth against all of these failure modes.

54. This is the vulnerability we consider most important for the Committee to understand because it creates dangerous incentives that could amplify harm during a crisis.

Current framework (neither capability assured):

55. ML system shows anomalous behaviour ⇒ Operator wants to disable ⇒ Technical capability exists? ⇒ [No assurance] ⇒ Can service survive shutdown? ⇒ [No assurance] ⇒ Perverse incentive: keep running ⇒ Harm accumulates / anomaly persists / behaviour escalates

With proposed amendment (both capabilities required):

56. ML system shows anomalous behaviour ⇒ Operator wants to disable ⇒ Technical capability exists? ⇒ YES (planned) ⇒ Can basic service survive? ⇒ YES (degraded mode available) ⇒ Safe shutdown executed ⇒ Controlled service reduction ⇒ Harm contained

Background

57. During the Bill’s second reading, Chi Onwurah MP raised concerns about "the role of foreign technology in our supply chains, particularly kill switches," noting "questions have been raised on numerous occasions on the Floor of the House about the prevalence of kill switches in Chinese technology."[27] This remark highlights what happens if critical systems become unavailable due to foreign technology dependencies, and this concern applies with particular force to autonomous or adaptive ML systems in CNI, which may become deeply embedded in

[27] Hansard, HC Deb, 6 January 2026, Cyber Security and Resilience Bill Second Reading. Chi Onwurah MP. https://hansard.parliament.uk/commons/2026-01-06/debates/BB815F91-651E-4A24-AAFE-8BD7D92B2033/CyberSecurityAndResilience(NetworkAndInformationSystems)Bill

operational processes and may be difficult to substitute quickly given their learned behaviours and market concentration. Fallback capability is key to the resilience of UK CNI systems in the face of this threat model. If a system becomes unavailable, whether through operator decision, system failure, or external circumstances, essential functions must continue.

58. Ben Lake MP raised the question of emergency shutdown powers for AI systems in critical infrastructure.[28] The CLTR report *Preparing for AI security incidents* argues that the UK Government "does not have the necessary powers to intervene in a crisis" and specifically recommends emergency powers to "contain an incident by temporarily reducing or removing public access to an AI model."[29]

59. The *International AI Safety Report October 2025 Key Update* recommends "kill-switch/corrigibility measures before deploying agentic systems to high-stakes environments"[30] (corrigibility research addresses the issue of systems becoming harder to reliably shut down as they become more capable). The recommendation issued in the Report Update reflects consensus among over 100 AI experts from 30 nations that this concern warrants proactive measures. The concern is grounded in evidence: as mentioned in a previous section, AISI evaluations show self-replication success rates rising from under 5% in 2023 to over 60% in 2025[31]. A system that can copy itself to new computational environments may not have a single point at which shutdown is effective.

60. The Bill already recognises the importance of reliable service provision to the extent that Regulation 14H(3)-(4) addresses supply chain substitutability, i.e. whether an operator can source equivalent services from alternative providers. This is necessary but not sufficient. Substitutability ensures procurement from a different provider, whereas operational continuity asks: "can this hospital, grid, or water system function at all if a core autonomous or adaptive system shuts down, regardless of why it shut down?”

On Section 43 (Directions to regulated persons):

61. The Bill contains a relevant provision, namely shutdown authority under specific exceptional circumstances. Section 43(3)(f) empowers the Secretary of State to direct "removing, disabling or modifying goods or facilities" where a security compromise poses risk to national security. However, this provision does not address the gaps we identify:

- **Assumption of a technical capability that may not exist.** Section 43 directs regulated persons to disable systems, but nothing requires this capability to exist before a crisis occurs. Metaphorically speaking, the Bill creates authority to press a button without ensuring the button exists or works. As autonomous systems become more

[28] Hansard, HC Deb, 6 January 2026, Cyber Security and Resilience Bill Second Reading. Ben Lake MP. https://hansard.parliament.uk/commons/2026-01-06/debates/BB815F91-651E-4A24-AAFE-8BD7D92B2033/CyberSecurityAndResilience(NetworkAndInformationSystems)Bill

[29] Tommy Shaffer Shane, Preparing for AI security incidents: Improving emergency preparedness with the UK AI bill and beyond, Centre for Long-Term Resilience, September 2025. https://www.longtermresilience.org/reports/preparing-for-ai-security-incidents/

[30] Yoshua Bengio et al, International AI Safety Report 2025: First Key Update, October 2025. https://internationalaisafetyreport.org/

[31] AI Security Institute, Frontier AI Trends Report, December 2025, p. 30, Figure 16. https://www.aisi.gov.uk/frontier-ai-trends-report

agentic, maintaining effective shutdown capability requires active engineering effort.

- **A reactive rather than preventive approach.** Section 43 is a direction power exercisable after a problem is identified. It does not require operators to maintain shutdown capability in advance. When a crisis emerges, the technical capability to comply may not exist, and building it under pressure may not be feasible.

- **Too slow for anomalies in autonomous or adaptive ML systems.** Section 43(9) requires consultation with the regulated person "so far as it is reasonably practicable." The consultation exemption (s43(10)) applies only where consultation would be contrary to the interests of national security, not where speed is the concern. Autonomous systems operating at machine speed may cause significant harm in the time consultation may take.

- **No fallback requirement.** Section 43 can direct shutdown but provides no assurance that shutdown is survivable. If complying with a shutdown direction causes unmitigated service blackout, operators of critical infrastructure face difficult choices, especially if their systems are high-level CNI hubs.

Our amendment addresses both gaps, requiring technical shutdown capability to exist before crises occur, and ensuring fallback so that shutdown, however triggered, does not cause catastrophic full blackouts.

62. The CrowdStrike incident of July 2024 demonstrated what happens when widely deployed software fails simultaneously across CNI without fallback:

- A single faulty update crashed approximately 8.5 million systems worldwide.[32]
- Airlines, hospitals, emergency services, banks, and government agencies were affected simultaneously.
- Estimated damages exceeded $5 billion for Fortune 500 companies alone.[33]
- Oxford Blavatnik School of Government professor and former head of the National Cyber Security Centre Ciaran Martin commented: "This is a very, very uncomfortable illustration of the fragility of the world's core internet infrastructure."[34]

63. The CrowdStrike incident was not AI-related. It was a deterministic failure from a code error rather than the consequence of unpredictable behaviour. We cite it to demonstrate the structural vulnerability whereby systems without fallback capability cause widespread disruption when disabled, regardless of what triggers the disabling. The fallback requirement ensures operators do not become so dependent on a single system that disabling it causes full service blackout, and instead plan for a backup degraded mode.

64. This failure mode compounds with autonomous/adaptive ML systems:

[32] CISA, "Widespread IT Outage Due to CrowdStrike Update," July 2024. https://www.cisa.gov/news-events/alerts/2024/07/19/widespread-it-outage-due-crowdstrike-update

[33] Parametrix analysis, 2024. https://www.parametrixinsurance.com/reports-white-papers/crowdstrikes-impact-on-the-fortune-500

[34] Ciaran Martin, quoted in Financial Times coverage of CrowdStrike incident, July 2024. https://www.ft.com/content/fba9b61d-efcf-4348-b640-ccb1f9d18ced

	CrowdStrike	Autonomous/adaptive ML system in CNI
Cause	Clear bug in code	Outputs emerge from learned patterns
Diagnosis	Engineers trace explicit logic	No explicit logic to trace
Recovery	Normal operation pattern fully known Fix deployed, systems recovered	"Normal operation" may itself drift unpredictably
Speed	Static once broken	Autonomous decisions compound faster than human review

65. What compounds the problem in the context of autonomous/adaptive ML systems is the combination of autonomy, opacity, and speed: the system does things we did not explicitly specify; we cannot easily trace why; and it acts faster than we can check. Fallback capability is more urgent when the system requiring shutdown is unpredictable.

Recommendations

66. To ensure that safety decisions are not constrained by dependency or technical lacunae, we recommend requiring the maintenance of both capabilities: (a) the technical ability to disable or isolate autonomous or adaptive ML systems, with assurance that this capability will remain effective as systems become more agentic; and (b) the capacity to continue to deliver a basic level of essential functions during and after disabling, whether initiated by the operator, directed by authority, or caused by system failure or other circumstances. This follows the logic of fire safety: the time to ensure exits are clear is before, not during, the fire.

67. We recognise the Committee may wish to consider a more graduated approach. If so, we would suggest at minimum requiring the Secretary of State to publish guidance on operational continuity and shut down capability within 12 months of commencement, with a duty to make regulations if the guidance proves insufficient.

PROPOSED AMENDMENT – Section 29

68. Add new subsection:

> *"(X) Regulations under this Part must require persons carrying on an essential activity to—*
>
> > *(a) identify which functions are essential to continued provision of the service;*
> >
> > *(b) maintain the capability to operate those functions in a degraded mode should any network and information system incorporating autonomous or*

> *adaptive capabilities based on machine learning become unavailable or require isolation;*
>
> *(c) maintain the technical capability to disable or isolate any such network and information system—*
>
> > *(i) at the operator's discretion when continued operation poses risk to security or safety, or*
> >
> > *(ii) when directed by the relevant competent authority for reasons relating to the security or safety of the network and information system or the services it supports;*
>
> *(d) ensure that essential functions can continue in degraded mode during and after the exercise of the capability in paragraph (c), or following any other event that renders such a network and information system unavailable;*
>
> *(e) periodically test the capabilities in paragraphs (b) and (c) and report the results to the competent authority."*

69. Note on interpretation: Paragraph (a) requires operators to determine in advance which functions must continue; this scoping exercise is a prerequisite to meaningful fallback planning. Paragraph (c) addresses the technical shutdown capability concern: as autonomous/adaptive ML systems become more capable of self-replication and distributed operation, maintaining effective shutdown capability requires active engineering effort. Paragraph (c)(i) enables operators to act on their own judgment when they identify a risk. Paragraph (d) covers "any other event," ensuring fallback serves as defensive depth regardless of why the system becomes unavailable. The requirement is not that disabling causes no disruption, but that essential functions remain available throughout.

70. Alternative: bolster the designation criteria in Regulation 14H(3) - note that this alternative addresses fallback capability in OES, but does not address technical shutdown ability:

> *"(3A) In reaching a conclusion for the purposes of paragraph (1)(c), a designated competent authority must also have regard to whether the person has the capability to maintain essential operations in a degraded mode should supply be disrupted."*

5. No baseline requirements for designated critical suppliers

71. **VULNERABILITY:** designation power exists without corresponding obligations, leaving critical supply chain concentration risks unaddressed by default.

72. **THREAT MODEL:** a provider of ML-based services is designated as a critical supplier. Designation carries no specific obligations. Without baseline requirements applying to all designated suppliers, competitive pressure discourages investment in security practices that buyers cannot easily evaluate. The supplier's security practices remain opaque. A vulnerability in the supplier's systems propagates to multiple CNI operators simultaneously, causing distributed failures, potentially across several infrastructure chains.

Background

73. Section 12 creates a power to designate critical suppliers. As drafted, the Bill does not impose specific obligations on designated suppliers. The Bill creates a power to define such duties in secondary legislation, but does not require their creation. This means designation could be empty of meaning in practice, leaving vital parts of UK CNI exposed.

74. As noted above (paragraph 47), the *International AI Safety Report* identifies market concentration as a systemic risk.

Recommendations

75. We recommend establishing baseline expectations for designated critical suppliers, ensuring designation carries meaningful obligations and builds robustness from the outset.

PROPOSED AMENDMENT – Section 12

76. Add new subsection:

> *"(X) Regulations under this section designating a person as a critical supplier must specify—*
>
> > *(a) security measures the supplier must implement, which may be specified by reference to the Cyber Assessment Framework or a code of practice under section 36;*
> >
> > *(b) information the supplier must provide to the relevant competent authority regarding dependencies in its supply chain;*
> >
> > *(c) incident reporting obligations equivalent to those applicable to operators of essential services."*

6. Public sector cyber resilience reporting

77. **VULNERABILITY:** the public sector is excluded from the Bill's scope despite being a major cyber security target. Parliament has no systematic visibility of public sector cyber resilience.

Background

78. The Government Cyber Action Plan acknowledges that "historical underinvestment in both technology estates and proportionate cyber security measures have left us with a significant technical debt."[35]

79. The Synnovis attack of June 2024 delayed over 11,000 NHS appointments and, according to a written Government statement, "tragically, contributed to the death of a patient."[36] This attack targeted a private sector supplier, demonstrating how public services can be disrupted through supply chain vulnerabilities. Public authorities directly operating network and information systems face similar risks but without equivalent reporting requirements.

80. Julia Lopez MP noted during second reading that "the new obligations in this Bill broadly do not touch the public sector, where cyber-risk remains red-light-flashingly large."[37] Victoria Collins MP observed that "Government institutions and councils will still lack statutory protections."[38]

Recommendations

81. We do not propose extending the Bill's regulatory scope to public authorities, recognising this would be a significant expansion. Nonetheless, the existing reporting requirement in Section 40 could be amended to require assessment of public sector cyber resilience, including particular scrutiny over autonomous or adaptive ML systems. This would generate parliamentary intelligence and accountability without regulatory expansion. Accordingly, we recommend amending Section 40 to require assessment of public sector cyber resilience.

PROPOSED AMENDMENT – Section 40

82. After Section 40(2), insert:

> *"(2A) A report under this section must include—*
>
> > *(a) assessment of the cyber security and resilience of public authorities providing essential public functions;*

[35] Government Cyber Action Plan, January 2025. https://www.gov.uk/government/publications/government-cyber-action-plan

[36] Written Statement UIN HCWS1221, 6 January 2026. Ian Murray MP. https://questions-statements.parliament.uk/written-statements/detail/2026-01-06/hcws1221

[37] Hansard, HC Deb, 6 January 2026, Cyber Security and Resilience Bill Second Reading. Julia Lopez MP. https://hansard.parliament.uk/commons/2026-01-06/debates/BB815F91-651E-4A24-AAFE-8BD7D92B2033/CyberSecurityAndResilience(NetworkAndInformationSystems)Bill

[38] Hansard, HC Deb, 6 January 2026, Cyber Security and Resilience Bill Second Reading. Victoria Collins MP. https://hansard.parliament.uk/commons/2026-01-06/debates/BB815F91-651E-4A24-AAFE-8BD7D92B2033/CyberSecurityAndResilience(NetworkAndInformationSystems)Bill

(b) assessment of risks arising from public authority use of autonomous or adaptive systems based on machine learning;

(c) the Secretary of State's view on whether extension of requirements under this Act to public authorities would be appropriate, and if not, why existing measures are sufficient."

Conclusion

83. The vulnerabilities we have described arise from properties of autonomous and adaptive ML systems that make their behaviour difficult to predict, audit, and control, all the more so as their capabilities and the reliance placed upon them increase and become more interconnected across CNI chains. The Bill's existing framework is well designed for conventional cybersecurity threats but does not fully address these distinctive properties.

84. The amendments we propose are designed to:

- Enable systematic learning about novel failure modes before catastrophic incidents occur.
- Match reporting and review cycles to the pace of capability change.
- Complete the Government's existing work on AI cyber security guidance by creating pathways to enforceable standards.
- Ensure that autonomous systems in critical infrastructure can be safely shut down and that fallback capability bounds damage from any shutdown trigger, preventing lock-in and perverse incentives that could amplify harm.
- Address structural dependency and concentration risks that could propagate single-point failures across CNI.
- Create parliamentary accountability for public sector cyber resilience.

85. The Bill's purpose is sound. These amendments address gaps that become visible when examining how autonomous and adaptive ML systems actually behave, and serve to future-proof this critical piece of legislation.

Appendix

Incident intelligence for autonomous and adaptive ML systems - Full amendment text

– Amendment Part 1: Section 15 (Incident significance criteria)

- Part 2, Chapter 2(3) amending Regulation 11(4), page 22, line 16: add a subclause 11(4)(f) as follows:

  > *“(f) whether the incident involves failure modes not previously observed in the relevant sector materially involving autonomous or adaptive systems based on machine learning, including where the potential impact of such failure modes was mitigated or prevented.”*

- Part 2, Chapter 2(3) amending Regulation 11(5), page 22, line 26: add a subclause 11(5)(f) as follows (current subclause (f) becomes subclause (g)):

  > *“(f) where the incident was associated with one or more autonomous or adaptive systems based on machine learning, details of those systems and their involvement in the incident;”*

- Part 2, Chapter 2(3) introducing Regulation 11A, page 23 line 13 to page 24 line 6, replace clauses (3-7) with the following clauses (3-8):

  > *“(3) For the purposes of this regulation, an incident is a “data centre incident” if —*
  >
  > *(a) the incident has affected or is affecting the operation or security of the network and information systems relied on to provide the data centre service provided by the OES, and*
  >
  > *(b) the impact of the incident in the United Kingdom or any part of it has been, is or is likely to be significant having regard to the factors listed in paragraph (4).*
  >
  > *(4) The factors referred to in paragraph (3)(b) are—*
  >
  > *(a) the extent of any disruption which has occurred, is occurring or is likely to occur in relation to the provision of the essential service provided by the OES;*

(b) the number of users which have been affected, are being affected or are likely to be affected;

(c) the duration of the incident;

(d) the geographical area which has been affected, is being affected or is likely to be affected by the incident;

(e) whether the confidentiality, authenticity, integrity or availability of data relating to users of the essential service has been, is being or is likely to be compromised;

(f) whether the incident involves failure modes not previously observed in the relevant sector materially involving autonomous or adaptive systems based on machine learning, including where the potential impact of such failure modes was mitigated or prevented.

(5) The information referred to in paragraph (2)(b) is—

(a) the OES's name and the data centre service to which the incident relates;

(b) the time the incident occurred, its duration and whether it is ongoing;

(c) information concerning the nature of the incident;

(d) where the incident was caused by a separate incident affecting another regulated person, details of that separate incident and of the regulated person in question;

(e) information concerning the impact (including any cross-border impact) which the incident could have had, has had, is having or is likely to have (as the case may be);

(f) where the incident involved one or more autonomous or adaptive systems based on machine learning, details of those systems and their involvement in the incident.

(g) such other information as the OES considers may assist the designated competent authority in exercising its functions under regulation 11B in relation to the incident.

(6) The notifications required by paragraph (2) must be given—

(a) in the case of an initial notification, before the end of the period of 24 hours beginning with the time at which the OES is first aware that a data centre incident has occurred or is occurring;

(b) in the case of a full notification, before the end of the period of 72 hours beginning with that time.

(7) A notification under paragraph (2) must be in writing, and must be

provided in such form and manner as the designated competent authority determines.

(8) An OES must send a copy of a notification under paragraph (2) to the CSIRT at the same time as sending the notification to the designated competent authority for the OES."

- Part 2, Chapter 2(3) introducing Regulation 12(A)(3), page 26, line 38: add a subclause 12A(3)(h) as follows:

  "(h) whether the incident involves failure modes not previously observed in the relevant sector materially involving autonomous or adaptive systems based on machine learning, including where the potential impact of such failure modes was mitigated or prevented."

- Part 2, Chapter 2(3) introducing Regulation 12(A)(4), page 27, line 8: add a subclause 12A(4)(f) as follows (current subclause (f) becomes subclause (g)):

  "(f) where the incident was associated with one or more autonomous or adaptive systems based on machine learning, details of those systems and their involvement in the incident;"

- Part 2, Chapter 2(3) introducing Regulation 14E(3), page 30, line 9: add a subclause 14E(3)(f) as follows (current subclauses (f) and (g) becomes subclauses (g) and (h)):

  "(f) whether the incident involves failure modes not previously observed in the relevant sector materially involving autonomous or adaptive systems based on machine learning, including where the potential impact of such failure modes was mitigated or prevented;"

- Part 2, Chapter 2(3) introducing Regulation 14E(4), page 30, line 22: add a subclause 14E(4)(f) as follows (current subclause (f) becomes subclause (g)):

  "(f) where the incident was associated with one or more autonomous or adaptive systems based on machine learning, details of those systems and their involvement in the incident;"

– Amendment Part 2: Section 18 (Information sharing)

- Part 2, Chapter 3, section 18(3) substituting Regulation 6, page 41, line 20, insert a clause 6C as follows:

> *“**Sharing of information with the AI Security Institute***
>
> ***6C.**—(1) Where the CSIRT receives notification of an incident under regulation 11, 11A, 12A, or 14E that materially involves autonomous or adaptive systems based on machine learning, the CSIRT must share relevant technical information with the AI Security Institute within 72 hours.*
>
> *(2) In this regulation, 'AI Security Institute' means such body as the Secretary of State may designate for the purposes of this regulation.”*

January 2026