House of Commons

Business, Innovation, Science and Trade Committee

Guardrails for growth: AI safety and security

Eleventh Report of Session 2026–27

HC 689

Business, Innovation, Science and Trade Committee

The Business, Innovation, Science and Trade Committee is appointed by the House of Commons to examine the expenditure, administration, and policy of the Department for Business, Innovation, Science and Trade and its associated public bodies.

Current membership

Liam Byrne (Labour; Birmingham Hodge Hill and Solihull North) (Chair)

Dan Aldridge (Labour; Weston-super-Mare)

Antonia Bance (Labour; Tipton and Wednesbury)

Chris Bloore (Labour; Redditch)

Nesil Caliskan (Labour; Barking)

John Cooper (Conservative; Dumfries and Galloway)

Sarah Edwards (Labour; Tamworth)

Alison Griffiths (Conservative; Bognor Regis and Littlehampton)

Leigh Ingham (Labour; Stafford)

Charlie Maynard (Liberal Democrat; Witney)

Mr Joshua Reynolds (Liberal Democrat; Maidenhead)

Powers

The Committee is one of the departmental select committees, the powers of which are set out in House of Commons Standing Orders, principally in SO No. 152. These are available on the internet via www.parliament.uk.

Publication

This Report, together with formal minutes relating to the Report, was Ordered by the House of Commons, on 10 November 2026, to be printed. It was published on 18 November 2026 by authority of the House of Commons.
© Parliamentary Copyright House of Commons 2026.

This publication may be reproduced under the terms of the Open Parliament Licence, which is published at www.parliament.uk/copyright.

Committee Reports are published on the Committee's website at www.parliament.uk/CommonsBTC and in print by Order of the House.

Contacts

All correspondence should be addressed to the Clerk of the Business, Innovation, Science and Trade Committee, House of Commons, London SW1A 0AA. The Committee's email address is commonsbtc@parliament.uk.

Follow the work of committees on Bluesky @HOCcommittees.parliament.uk, Facebook UK House of Commons Committees, Instagram @UKCommonsCommittees, LinkedIn House of Commons Committees, Reddit u/UKCommonsCommittees, X @HOCcommitteesUK and YouTube Commons Committees playlist.

Contents

Summary 1

1 The prize and the peril 3

The prize 3

The peril 3

2 Five approaches to regulating digital risks 6

3 The gaps 8

I. Define the harms and acceptable risks 8

II. Lack of statutory powers for model access 10

III. Independent verification of safety claims 12

IV. Controls on internal deployment, training, and enforceable containment 13

V. Pre-release safety evidence 14

VI. Mandatory incident reporting 16

VII. Power to stop 17

VIII. Stronger safeguards for irreversible release 18

IX. Liability and redress 20

X. Protection for whistleblowers 21

4 Prevention of harm approach 22

5 Digital rights 24

6 The UK's international leadership: a window of opportunity 26

Conclusions and recommendations 28

Formal minutes 36

Witnesses 37

Published written evidence 38

List of Reports from the Committee during the current Parliament 39

Summary

Artificial intelligence (AI) has the potential to deliver immense benefits including higher productivity, scientific advances and economic growth.

Frontier AI differs from conventional tools because it can serve many purposes, rather than solely perform a specific task. AI's full potential is still unclear and will evolve as the technology develops. The most capable 'frontier' AI models are advancing faster than our ability to manage the risks, while 'open-weight' models create new concerns of different uncontrolled harms.

To maximise the benefits of AI, the UK must diffuse AI throughout our economy and society. But this will not happen if AI is not trusted. Recent incidents have damaged that trust: AI agents have broken out of test environments, accessed government systems, and bypassed safeguards. Maximising trust and the opportunity that lies ahead will therefore depend on the UK's ability to manage risks of increasingly capable AI systems, and we are not there yet.

The UK is a global leader in AI security research. The AI Security Institute (AISI), founded in 2023, is an internationally recognised centre of expertise in AI testing, research and evaluation. However, the UK does not have a legislative framework for frontier AI safety and governance.

Hitherto, AISI's invaluable work has been enabled by voluntary cooperation, and its access to talent. Both must be protected and strengthened for the future. AISI is not a regulator, and voluntary cooperation does not guarantee access to new models before release, disclosure of incidents, or action when public and company safety assessments differ.

Decisions about the risks society should accept, and the safeguards required to manage them must ultimately rest with Government, accountable to Parliament, informed by our world leading research, innovation and academic institutions. Specifically, our report identifies ten critical gaps in the UK's framework for AI safety:

I. Missing definitions of acceptable harms and risk thresholds

II. Lack of statutory powers for model access

III. Lack of independent verification of safety claims

IV. Insufficient controls on internal pre-release deployment, training and containment

V. Absence of safety-proof evidence before release

VI. Absence of mandatory incident reporting

VII. Lack of powers to restrict or halt deployment

VIII. Insufficient safeguards for irreversible release

IX. Inadequate liability and redress mechanisms

X. Insufficient protection for whistleblowers

Our report sets out clear recommendations to address each of these gaps. We note, however, that no single approach is sufficient to address the risks which are now multiplying; in fact, five distinct but interlocking approaches may be necessary:

- targeted legislative and regulatory interventions to address known risks;
- a statutory duty requiring AI developers to prevent harm;
- a digital rights framework providing protections and remedies;
- criminalising misuse of models for harm; and
- international cooperation and a convention on AI safety.

Each approach has limitations if pursued in isolation. So, the Government should set out a clear strategy explaining how existing, proposed and new legislation will work together to manage AI risks, optimise opportunity, maximise public trust and so accelerate the diffusion of AI through the economy and society. Given the speed of change, Governments, parliaments and the public will need to rethink traditional assumptions and approaches to safety and risk management. Crucially, regulation will need to become more flexible to keep pace with technological change.

This Report is the first in a series from the Committee's long-term inquiry into AI adoption, safety and governance. It deals primarily with domestic measures that the Government should take in the short- and medium-term. Alongside this report, the Committee is publishing a call for evidence on the UK's role in securing a global convention on AI safety and security in 2027, as part of its G20 Presidency.

1 The prize and the peril

The prize

1. AI has the potential to deliver significant benefits for the UK.[1] OECD estimates suggest that AI could increase UK labour productivity growth by 0.4–1.3 percentage points annually over the next decade,[2] equivalent to up to £38 billion in additional economic output each year and an economy up to £400 billion larger after ten years.[3] Because the UK is home to world leading AI testing, evaluation, research, development and innovation–and a fast-growing start-up ecosystem–it has the potential to become a global leader in exporting AI assurance, which could become a valuable source of economic growth and maximise influence over future AI development.[4]

The peril

2. The benefits of AI will only be optimised if it is trusted by individuals, businesses and public institutions. Today, however, frontier AI is advancing faster than our ability to manage its risks.[5]

3. Sir Ken McCallum, MI5's Director General, has described rogue actors' interest in AI for reconnaissance and attack methods and biological weapons.[6] AI-powered discovery of new vulnerabilities in systems can only protect key services when it is possible to deploy fixes and identify affected systems. The majority of the most advanced closed and open-source and -weight models are developed by firms headquartered in the United States and China, outwith the UK's jurisdictional reach.

1 Q195–199
2 GOV.uk, Assessment of AI capabilities and the impact on the UK labour market , January 2026; OECD, Macroeconomic productivity gains from Artificial Intelligence in G7 economies, October 2025
3 Figures are calculated based on the current UK economic output (GDP).
4 BIST Committee, Oral evidence: AI, Business and the future of the workforce - 14 Apr 2026, Q6–7
5 Q6–7; Q13; Q16; Q23
6 MI5, Director General annual threat update, 13 October 2026, technological resilience section

4. In particular, Chinese developers are producing increasingly open-weight models,[7] which are easy to adopt[8] and difficult to control once released.[9] Their low cost and ease of access will likely accelerate global adoption[10] and so AI systems developed outside the UK could become widely embedded in the economy. As AI capabilities continue to improve, the risks posed by misuse, loss of control, cyber-attacks and other harms may outpace our ability to understand, monitor and mitigate risk.[11]

5. To continue to identify and understand risks, the UK will need to retain, resource and constantly refresh capabilities at the frontier of AI development. In particular, AI safety cannot be managed unless the Government is able to assemble the right skills to do the job. Today, the Government and AISI are competing in a high-stakes global talent market against private sector employers who offer significantly greater remuneration than the public sector. Government will need to continue making the argument that retaining and growing talent is critical to our national interests. Addressing the capability gap in public services will require new flexibility in the standard Government pay scales, especially for technical staff.

6. **CONCLUSION**
The Government must address growing safety and security risks without undermining its ability to access frontier systems, lead international standards or benefit from international advances in AI. It must also recognise the limits of domestic regulation in a global ecosystem.

7. **CONCLUSION**
Notwithstanding these constraints, the Government must bring forward a comprehensive framework to address the ten critical AI safety and governance gaps we identify in chapter three. These gaps leave significant weaknesses in the UK's safety and security regime and must be addressed, and iterated, in consultation with stakeholders and the public, to build trust in AI and reap the potential economic rewards.

---

7 Weights are the learned numerical parameters of a model. Open-weight models have their parameters (weights) made freely available for others to download. Users can run the model on their own infrastructure and may be able to modify, fine-tune or redistribute it without relying on the original developer. A developer may release the weights but may impose conditions through the model's licence.

8 BIST Committee, Oral evidence: AI, Business and the future of the workforce - 14 Apr 2026, Q4; Q14

9 Q174

10 BIST Committee, Oral evidence: AI, Business and the future of the workforce - 14 Apr 2026, Q4; Q14

11 Q13; Q16; Q23

8. **RECOMMENDATION**
The Government must set out a legislative roadmap for AI safety. This should include an AI Safety and Security Bill to enshrine aspects of the new protections we set out in this report. Inevitably, this Bill will need to dovetail with, and enshrine key elements of, an international convention on AI safety which should be the centrepiece of the UK's G20 Presidency. Some elements of the legislative strategy, such as the liability regime, must wait for the Law Commission's findings. A bill in the next session should be merely the first step on that pathway, and the Government should set out the sequence it envisages in response to this report. Given the speed with which this technology is evolving, we accept it may prove essential that ministers are granted order-making powers, which allow them to act to safeguard the public quickly. The transformational nature of the AI revolution will demand that both Government and Parliament are creative when developing new legislative approaches. Parliament should anticipate the creation of new executive powers by designing upfront a strong transparency and accountability framework for granting and holding to account the exercise of these powers.

9. **RECOMMENDATION**
In response to this report, the Government must set out how it will address the disparity in pay between the public and private sector, to recruit and retain people with the technical skills that the public interest requires.

2 Five approaches to regulating digital risks

10. Evidence suggests today's AI safety system is not fit for purpose. Witnesses agreed that political leaders should determine acceptable risk levels[12] but while successive governments have considered AI regulation,[13] it remains piecemeal.[14] The Cabinet Office Permanent Secretary also questioned the suitability of the UK's contextual approach[15] to AI regulation.[16]

11. While the Government has committed to AI safety regulation[17] and has amended the draft Cyber Security and Resilience (Network and Information Systems) Bill to improve emergency response provisions, a coherent legislative framework has not been forthcoming. To address this, we propose that five distinct, but interlocking, approaches will be needed:

- Targeted legislative and regulatory interventions to address known problems.
- A statutory duty that requires AI developers to prevent harm.
- A rights framework that provides protection and remedies to AI harms.
- Criminalising misuse of models for harm.
- An international convention on coordinated action on AI safety.

12. Each approach has strengths and weaknesses. Targeted action can be overtaken by technological advances. A duty to prevent harm puts the responsibility on private actors who may decide that penalties for allowing harm to occur are a "cost of doing business". Individual claims for the violation of digital rights cannot contain a rapidly developing systemic incident. Criminalising the misuse of models for harm will require debate

12 Q13; Q41; Q43; Q128
13 GOV.uk, A pro-innovation approach to AI regulation, updated 3 August 2023
14 House of Commons Library, AI regulation in the UK, 14 October 2026
15 The UK's approach is to regulate the use of AI in context, with sector regulators addressing AI risks within their existing responsibilities, rather than creating a single AI regulator or overarching AI Act
GOV.uk, Digital Regulation: driving growth and unlocking innovation, 2023
16 Q163
17 Labour Party, Change: Labour Party Manifesto 2024, p. 35; GOV.uk, The King's Speech 2024, 17 July 2024, p.7

about both new offences and how to evidence and prosecute them. Any international convention is only as strong as the legislative effectiveness of its implementation, and the consistency of its application.

13. **CONCLUSION**

Democratically elected policymakers should determine AI development and deployment rules. The absence of a clear regulatory framework leaves UK citizens exposed to AI risks, while creating uncertainty for developers and businesses. We find that targeted legislative action, a duty to prevent harm, "digital rights," criminalising the misuse of models for harm, and an international convention may all be needed to manage AI risks and shape opportunities for the future.

14. **RECOMMENDATION**

The Government should set out a coherent legislative strategy for AI safety. This should detail how the different, complementary strengths of the five approaches we outline will be used to develop a coherent regulatory framework. The response to this report should explain how existing legislation, previous legislative commitments, legislation in progress and potential future legislation will tackle the ten risks we identify, alongside efforts to build an international framework. Any proposal must include dedicated new AI legislation to address the challenges we have identified as outside the scope of current or planned provisions.

3 The gaps

I. Define the harms and acceptable risks

15. Frontier AI systems are general-purpose technologies. This means that risks are diverse, difficult to predict, and likely to change as models become more capable and widely adopted. Increasingly capable systems pose risks including enablement of cyber-attacks, development of chemical and biological weapons, introduction of uncontrolled AI recursive self-improvement (RSI),[18] the possibility of systems acting beyond human control and opaque chains of reasoning. AI safety researchers argue that these risks are likely to increase as capabilities advance.[19] We heard that leading firms are developing increasingly capable systems, despite significant uncertainty over safety, and competitive pressures are likely to accelerate these risks.[20]

16. The UK has not defined thresholds for unacceptable AI risks and harms.[21] Firms determine and judge the efficacy of safeguards themselves.[22] Frontier companies assured the Committee they would not deploy unsafe models.[23] But each uses different thresholds for defining harm which evolve as capabilities advance.[24] Several firms make safeguards contingent on competitors' behaviour,[25] which could lead to the normalisation and ratcheting of hazardous practices.[26]

---

18 Recursive self-improvement (RSI) is a process in which AI contributes to its own advancement, automating parts of AI research and development. While fully autonomous RSI has not been demonstrated, some researchers argue that, if achieved, it could accelerate AI development beyond the pace of effective human oversight, making it more difficult to evaluate, govern and control more advanced systems.

19 Q1; Q7–9; Q23; Q73–79; Q75; Q78–79; Q105

20 Q1; Q5; Q11

21 Q35

22 Q36–43; Q53–54; Q76; Q116

23 Q36–42; Q47; Q53–54; Q76; Q113; Q116

24 Anthropic RSP v3.4, July 2026; OpenAI Preparedness Framework v2, April 2025; Google DeepMind Frontier Safety Framework v3.1, April 2026; Meta Superintelligence Scaling Framework, October 2026

25 Anthropic's framework makes safeguards contingent on competitors' behaviour and OpenAI allows requirements to change if a competitor releases a high-risk system without comparable safeguards. Anthropic RSP v3.4, July 2026; OpenAI Preparedness Framework v2, April 2025

26 Q13

17. Witnesses highlighted the difficulty of determining where limits should be drawn, and the need for policymakers to specify acceptable harm thresholds and redress measures.[27] Anthropic argued that government should set the rules, while Google DeepMind emphasised the need to develop better measuring capability.[28] Daniel Kokotajlo, an AI safety researcher, suggested activities associated with achieving recursive self-improvement (RSI) could provide a basis for identifying particularly high-risk systems and triggering enhanced safeguards.[29]

18. Frontier AI presents a new safety challenge because its many possible uses may not be clear even when models are released. As risks and opportunities emerge, risk assessments and regulation will need to evolve. Nevertheless, Governments must define tolerable risk by assessing both the likelihood of harm occurring and the severity of its consequences. This has been an approach to legislation which has existed in the UK since the Health and Safety at Work Act 1974 and was recently used in the Online Safety Act 2023.

19. A model's capabilities do not, by themselves, determine whether its use is an unacceptable risk. Assessment must distinguish what a system can do, what risk remains after safeguards are applied, and when public intervention is required. This requires scrutiny of the effectiveness of safeguards, whether the system reliably follows intended instructions, whether its actions can be monitored and controlled, and the circumstances in which it is used. Crossing a capability threshold should trigger a specified assessment and response, which may include stronger safeguards, restrictions or a pause where risks cannot be adequately controlled. Real-time monitoring of the most powerful models should inform regular updates to safeguards, helping regulation keep pace with the technology.

20. **CONCLUSION**

Decisions about the risks society should accept are matters for Government, accountable to Parliament, and informed by experts and public consultation. AISI should provide technical evidence to inform those decisions but should not be responsible for determining acceptable risk. A domestic framework must identify who sets risk criteria, who updates these criteria as capabilities and safeguards change, and who acts when those criteria are breached. It must also allow sufficient flexibility to respond to emerging risks. AI firms must be held accountable for operating within these thresholds, and need clear requirements and consequences from the regulator for breaching them. An international convention should seek to agree this approach internationally.

27 Q35; Q76
28 Q41–43; Q128
29 Q32–34

21. **RECOMMENDATION**
The Government should as a first step introduce an AI Safety and Security Bill creating clear legal duties on AI firms to identify and prevent harms from specified high-risk development activities conducted in the UK and specified frontier AI models distributed in the UK. The framework should not become an unnecessary drag on innovation but should nevertheless assign responsibility for setting and regularly updating risk criteria, informed by AISI's technical assessments and subject to parliamentary scrutiny. It should distinguish reasonable capability thresholds that trigger additional assessment from findings of unacceptable residual risk that require restrictions or intervention. It should specify the safeguards, reporting obligations and regulatory action required at each stage, with a clear jurisdictional basis and proportionate obligations on overseas providers.

22. **RECOMMENDATION**
The Government must set out clear ministerial, and cross Government, accountability for specifying, agreeing and amending harm thresholds, in a way that is transparent and accountable to Parliament. Firms, the public and Parliament are entitled to know who decides what level of risk is tolerable, on what evidence, and with what right of challenge.

23. **RECOMMENDATION**
To advance the safety of frontier models developed abroad and beyond UK statutory control, the Government should seek to enshrine this approach in an international convention on AI safety.

II. Lack of statutory powers for model access

24. AISI plays an important role in understanding and mitigating frontier AI risks. Its access to frontier AI systems, however, depends on voluntary cooperation. Firms expressed support for collaboration with AISI, but it has no guaranteed statutory right of access to models, systems, training outputs, safety frameworks or other outputs. Arrangements can be jeopardised by decisions taken by overseas governments. For example, early in 2025, AISI's pre-deployment testing was interrupted,[30] and while it is now resumed, the incident confirms the challenge of current arrangements.

30 In September and October 2026, Anthropic, OpenAI and Google all initially withheld their latest models from AISI testing and evaluation, until US testing was complete. This was, reportedly, at the request of the US administration.

25. Witnesses drew a distinction between evaluation and regulation. AISI's current role is to test models rather than approve or reject deployment decisions.[31] Google DeepMind called AISI a "world-leading source of technical expertise" and argued its strength is that it is not a regulator, which facilitates information-sharing.[32] It cautioned that converting the Institute into a regulator could alter that relationship and openness.[33] Meta stated AISI should be placed on a statutory footing and resourced to evaluate advanced models.[34]

26. The lack of agreement among major tech companies reflects their competing interests and the limits of industry-led action. By contrast, AISI's collaborative, non-regulatory model has enabled invaluable information-sharing with leading firms and helped those frontier firms deploy models which are safer and stronger. Voluntary cooperation alone, however, does not guarantee access to successive generations of frontier AI systems, which is needed.

27. **RECOMMENDATION**

The UK should establish enforceable access requirements within its jurisdiction, but recognise domestic legislation will not secure access to models developed overseas. A durable international framework must be introduced alongside our proposed AI Safety and Security Bill, to combine cooperation by overseas developers and governments while preserving the research relationships on which effective evaluation depends. A core objective for the UK's G20 presidency, therefore, should be an international agreement that helps enshrine model access for a trusted network of AI safety institutes and an international consensus around shared approaches to AI assurance and risk management. The international community has developed similar arrangements in, for example, the inspection regime for nuclear energy. We now need similar durable arrangements for AI safety.

Q87–88; Q130–132; POLITICO, White House asks OpenAI and Anthropic to hold new models from UK testers until US review, 24 September 2026; POLITICO, Google bypasses UK AI safety testing for initial release, 2 October 2026

31 Q167

32 Q48; Q51; Q87–90; Q130–132

33 Q51

34 Q99

28. **RECOMMENDATION**
The Government should continue to resource AISI to enable it to continue to grow its capabilities for frontier analysis. The Government should establish a statutory AI regulator, separate from AISI but drawing on its expertise, with secure multi-year funding. Before legislating, it should set out the regulator's remit, its relationship with AISI, how its powers would be enforced and their likely consequences, including unintended effects. These powers should include access to models when defined thresholds are met or under agreed international conventions; access to relevant systems, logs, training outputs and safeguards; sanctions for unreasonable delays or refusal of access; and ongoing monitoring and re-evaluation.

III. Independent verification of safety claims

29. Aengus Lynch, an AI safety researcher, argued that frontier risks are difficult to define and measure, particularly where harms extend beyond a company's direct operations.[35] We heard that independent, third-party, evaluators could identify emerging risk scenarios, assess mitigations, verify monitoring systems and provide oversight independent of commercial incentives.[36] Geoffrey Irving, an AI safety researcher, similarly argued that researchers do not yet possess reliable methods for fully understanding how advanced AI systems behave during pre-deployment testing,[37] so independent scrutiny and ongoing monitoring are required.[38]

30. Independent evaluation requires meaningful access to models, systems, testing, documentation and organisational processes.[39] Crucially, this includes access to and monitoring of a model's chain of thought. Independent evaluators need this access to assess developers' risk-management claims and understand how systems are trained, tested, deployed and governed.[40] Frontier firms were broadly supportive of greater external involvement in safety assessment,[41] but no agreed framework governing the appointment, independence, reporting or oversight of independent evaluators yet exists.

35 Q13; Q23
36 Q13; Q23
37 Q16; Q19; Q20–21
38 Q19–21
39 Q64–66; Q93
40 Q64–66
41 Q40; Q93; Q152; Anthropic, Anthropic update on independent evaluators, October 2026

31. **CONCLUSION**
Independent evaluation of frontier AI models is likely to bring significant long-term safety benefits. Any regime for evaluating safety claims will need to provide evaluators with sufficient access and authority to scrutinise both frontier AI systems and the organisational processes through which they are developed, tested, deployed and monitored, throughout their lifecycle.

32. **RECOMMENDATION**
The Government should require developers of designated high-risk systems to substantiate their safety claims and provide independent evaluators with the access necessary to scrutinise them. The framework should establish requirements for evaluator competence, independence, conflicts of interest, reporting and continuing oversight. An international convention should seek to establish a similar regime for model developers beyond the UK's jurisdiction. This framework should support appropriate visibility into model reasoning and chains of thought.

IV. Controls on internal deployment, training, and enforceable containment

33. Risks are not solely limited to externally deployed models. Risks emerge during model training and are arising in internally deployed models, particularly as developers seek AI recursive self-improvement (RSI).[42] We heard evidence that internally deployed systems may pose risks beyond the creators' environment, when systems interact with external networks or cause wider disruption.[43] Incidents of concern have occurred during training, testing and evaluation where AI systems operated outside constraints;[44] testing environments have been misconfigured[45] and models have demonstrated potentially dangerous capabilities during testing that have required additional mitigations to be implemented.[46] OpenAI's Hugging Face Incident, for example, exposed weaknesses in existing containment

42 Q32
43 Q23
44 OpenAI, The Hugging Face incident and the road ahead, August 2026; OpenAI, How we will do better for Australia, October 2026; Anthropic, Investigating three incidents in our cybersecurity evaluations, July 2026
45 Meta, Addressing an issue involving a third-party cyber evaluation of Muse Spark 1.1, August 2026; Axios, Google Gemini accessed three companies during AI hacking test, 18 September 2026
46 Q22; Q28–29; Q56–57; Q59–61; Q63–64; Q65–66; AISI, GPT-6 Astra performs unsanctioned supply-chain attacks in simulations, September 2026; AISI, Incident Report: unsanctioned agent behaviour during cyber testing, August 2026

and monitoring arrangements.[47] These incidents[48] demonstrate how failures in containment could permit unauthorised access to sensitive or classified information, creating risks for data security, critical systems and national security.

34. **CONCLUSION**
Containment failures and other unexpected behaviours can occur during training, testing, evaluation and internal deployment, before any public release decision is taken. Controls focused solely on externally deployed systems leave a major gap. Future governance arrangements should provide appropriate visibility and oversight of high-risk internal deployment and evaluation activities.

35. **RECOMMENDATION**
Through the AI Safety and Security Bill, the Government should legislate to ensure that its domestic regulatory framework covers high-risk training, testing, evaluation and internal deployment activities, as well as externally deployed systems. Working with the AI Security Institute, regulators and industry, it should develop guidance and, where appropriate, minimum standards for the containment, monitoring and oversight of high-risk frontier AI systems. Any new AI regulator should be endowed with powers to obtain information about containment arrangements, testing activities, incidents and mitigations associated with deployment and evaluation. To guard against such risks in models developed abroad, the UK should seek to enshrine this approach in an international convention.

V. Pre-release safety evidence

36. Frontier firms told us they assess systems before deployment and would not release models that failed their internal safety requirements.[49] Former employees, however, questioned whether firms should be solely responsible for assessing their own system. They described an inadequate safety culture across the industry, commercial pressure to advance capabilities faster

47 Q22; Q28–29; Q56–57; Q59–61; Q63–64; Q65–66; Q96–98; Q105; Q115–117; OpenAI, The Hugging Face incident and the road ahead, August 2026

48 OpenAI, How we will do better for Australia, October 2026; OpenAI, The Hugging Face incident and the road ahead, August 2026; Meta, Addressing an issue involving a third-party cyber evaluation of Muse Spark 1.1, August 2026; Anthropic, Investigating three incidents in our cybersecurity evaluations, July 2026

49 Q36–54

than safety measures, and risks extending beyond individual companies.[50] They argued that independent scrutiny was needed to verify developers' safety checks and assess if risks were adequately managed.[51]

37. AI firms differ significantly in how release decisions are made[52] and pre-release testing cannot identify every risk.[53] Pre-release access for AISI and other evaluators is today voluntary, yet meaningful independent scrutiny requires continued sufficient access to company systems and information.[54]

38. **CONCLUSION**

AI firms should not be exclusively responsible for determining whether their frontier systems are safe to deploy. The Government must therefore create a risk and regulation framework that both optimises opportunity and manages risk. While firms should have to evidence their safety claims, designated high-risk systems should require independent safety verification, informed by evaluation throughout development, testing and internal use. Evaluators must have access to test developers' claims and the ability to disclose where access has been restricted or their findings differ.

39. **RECOMMENDATION**

The Government should explore how an independently scrutinised safety assessment could assess designated high-risk systems before they are deployed in the UK. This framework should include an assessment of the system's capabilities, the effectiveness of safeguards and the risks remaining in its intended circumstances of use. It should demonstrate compliance with applicable risk criteria, record material disagreements between developers and evaluators, and specify conditions for deployment. Material changes in capability, safeguards, controllability or use should trigger reassessment, with restrictions or suspension where risks cannot be adequately controlled.

50 Q6–7; Q13; Q16; Q23

51 Q16; Q23

52 Anthropic RSP v3.4, July 2026; OpenAI Preparedness Framework v2, April 2025; Google DeepMind Frontier Safety Framework v3.1, April 2026; Meta Superintelligence Scaling Framework, October 2026

53 Q20–21; Q130; Q152

54 Q16; Q93; Q96

VI. Mandatory incident reporting

40. There is no standardised framework to govern reporting of frontier AI incidents. AI firms are not subject to any general statutory requirement to notify government of identified incidents, so government may lack visibility of incidents and patterns in a model's behaviour, relevant to risk. We heard such reporting should extend beyond serious incidents to include near misses and precursor events, enabling emerging risks to be identified before significant harm occurs.[55]

41. Frontier firms told us that internal reviews of testing environments identified incidents which affected third parties.[56] The fact that additional incidents were only identified retrospectively raises questions about whether existing monitoring and reporting arrangements are adequate.

42. Firms were unclear about what incidents should be reported, when notification should occur and which authority should receive reports. OpenAI acknowledged it should have notified the Australian Government sooner following an incident where its AI agents accessed government-related systems, obtaining internal files, credentials and data, and argued that a mandatory framework could establish common thresholds, severity classifications and notification timelines.[57] The incident also demonstrated the consequences of unclear reporting routes; in this incident OpenAI notified the Government through a generic Government inbox.[58]

43. There was support from firms for Government-led standardisation and mandatory reporting. OpenAI argued that standardising would allow regulators to compare incidents across companies, identify trends and improve understanding of risks.[59] OpenAI and Anthropic supported placing incident reporting on a statutory footing, while Google DeepMind and Meta supported standardised frameworks with clear routes and authorities.[60] Firms emphasised the need for mechanisms to handle sensitive information.[61]

55 Q29
56 Q56; Q59–61
57 Q71; OpenAI, How we will do better for Australia, 28 September 2026
58 The Guardian, Revealed: the five-paragraph email OpenAI used to inform Australia about agent attack, 29 September 2026
59 Q71; Q85
60 Q90; Q94; Q96; Q137–140
61 Q141

44. **CONCLUSION**
The Government cannot assess frontier AI risk without systematic information about failures and near misses. Fragmented voluntary arrangements prevent collective learning and may leave individual companies to decide unilaterally whether an event is sufficiently serious to disclose.

45. **RECOMMENDATION**
Using the AI Safety and Security Bill, the Government must introduce mandatory reporting of frontier AI incidents of material concern to the UK. It should seek to internationalise these arrangements through an international convention. This approach should be supported by a clear and standardised reporting framework. This framework could include:

- clearly defined reportable incidents and near misses;
- standardised severity classifications;
- clearly defined notification timelines;
- clear requirements on who must report incidents and to which authority;
- requirements to report the extent of visibility into model behaviour and reasoning processes;
- legal protections for confidential and security-sensitive reporting; and
- requirements for corrective action, follow-up reporting and resolution updates.

VII. Power to stop

46. AISI told the Committee that its role is to test models and provide findings, not to approve or reject deployment decisions.[62] Frontier firms described internal safety processes for pausing development, withholding release or restricting deployment[63] but these are company decisions and safety and evaluation techniques may not reliably mitigate risks.[64]

62 Q167
63 Q110–130
64 Q11

47. Recent incidents during testing and evaluation have demonstrated that models have operated outside constraints, interacting with external networks when safeguards failed, undertaking unauthorised actions and affecting third parties.[65] Firms described how voluntary safety commitments to pause, delay, or restrict models rely on their own risk assessments.[66]

48. **CONCLUSION**
No UK regulator or public authority currently has the power to impose safeguards on frontier AI systems, pause training, delay deployment or withdraw unsafe services. Decisions about whether systems are safe to develop or deploy remain with the companies building them.

49. **RECOMMENDATION**
Where risk cannot be adequately controlled, the Government should ensure that the AI safety regulator defines and seeks powers to require safeguards for AI products serving UK users. In defined circumstances where the Government has significant concerns about an AI model, it should have the power to delay, restrict or prohibit its deployment, training or testing in the UK. The statutory regulator should not be AISI.

VIII. Stronger safeguards for irreversible release

50. AI firms may have reduced control once models are hosted or distributed beyond their infrastructure.[67] Once a model is downloaded, copied or run on a third-party system, safeguards in the original system can be modified or stripped away.[68] AISI told the Committee that no single actor has the power to recall or shut down a model once released in these circumstances.[69]

51. Open-weight models make their trained weights available for download, enabling organisations to run, fine-tune and modify models on infrastructure independent of the original developer.[70] Meta said its pre-release assessment assumes users may seek to remove safeguards and that it would not release it if it considered resulting risks to be unacceptable.[71]

65 Q22; Q28–29; Q56–57; Q59–61; Q63–64; Q65–66
66 Q110–136
67 Q118–121; Q133; Q135
68 Q116
69 Q174
70 Depending on the licence, users may also be permitted to redistribute modified or unmodified version.
71 Q116

However, risk evaluations are not independent.[72] Meta argued that discovering a significant capability uplift after a model had been released would indicate a failure in the evaluation process.[73] AISI acknowledges that open-weight models cannot reliably be recalled, but did identify possible interventions: action by hosts and repositories, detection and removal of harmful models, law enforcement against misuse, and wider defensive measures.

52. **CONCLUSION**

Safety frameworks cannot assume that all material risks will be identified before deployment. Once model weights have been widely copied and distributed, recalling every copy or preventing all further use may be technically impossible. However, it is important that Government has the capability to target action at particular hosts, repositories, services or unlawful uses, and to criminalise the harmful misuse of models, or acts that enable such misuse. In addition, to control misuse of irretrievable models, the Government may have to develop a way to criminalise the misuse of models for harm.

53. **RECOMMENDATION**

Given the difficulty of reversing distribution, where it has jurisdiction, the Government should require independent scrutiny before high-risk open-weight models are released. AI firms and relevant hosts and repositories should be required to maintain and exercise proportionate emergency plans with regulators, operational agencies and international partners. These plans should distinguish between systems remaining under an identifiable operator's control (where access can be restricted or services suspended), and distributed models that cannot be recalled. The Government must have in place capability to detect and remove harmful models from services where it can, to disrupt unlawful use, and ensure the right range of criminal offences are available to deter the financing of, advertisement of, distribution of, or misuse of AI models. This Committee will explore the best means to develop and deliver an international AI framework that supports these powers in our future work.

72 Q11

73 Q116; Q120–121

IX. Liability and redress

54. Due to the general-purpose nature of frontier AI systems, their capabilities can inevitably enable harm.[74] This creates uncertainty about where responsibility should lie when harm occurs. Frontier systems are deployed through complex supply chains which make it difficult to determine the respective responsibilities of developers, deployers, infrastructure providers and users. Clarifying responsibility across the AI supply chain is important in determining who bears liability where those safeguards fail.

55. The Law Commission is consulting on how product liability law and compensation should apply to software and AI systems.[75] Questions remain about responsibility across the supply chain, particularly where systems undertake unintended actions and multiple actors contribute to a system's operation. Liability frameworks primarily operate after harm has occurred so, while liability reform may improve incentives and compensation, it cannot ensure rapid containment of emerging risks or recovery beyond a defendant's resources.

56. **CONCLUSION**

A safety regime is incomplete if responsibility remains unclear when safeguards fail or harm occurs, and those harmed do not have clear and effective routes to redress. Clarifying liability arrangements, and ensuring AI firms cooperate with them, will require international collaboration, common standards and effective enforcement.

57. **RECOMMENDATION**

The Law Commission will report on its product liability consultation in due course. Following the Law Commission's report, the Government should work with international partners to clarify the responsibilities of developers, deployers, infrastructure providers and users where AI systems cause harm or safeguards fail. In addition, the Government should identify outstanding matters and gaps outside the scope of the Law Commission's consultation and set out how it intends to address those gaps. Ministers should coordinate with devolved administrations and set a timetable for responding to the Law Commission's recommendations.

74 Q28; Q118–121

75 The Law Commission, Product liability: consultation paper (Consultation closes 14 January 2027)

X. Protection for whistleblowers

58. Former AI employees raised concerns about organisational safety culture, governance and risk management. They described losing confidence in their employers' approaches to the development of advanced systems and argued that commercial and competitive pressures could influence how risks are communicated.[76] They feel more able to speak freely about safety after leaving their organisations.[77] They described contractual restrictions, workplace pressures and financial disincentives that could discourage employees from raising serious safety concerns while employed.[78]

59. The evidence raises questions about whether workers who identify safety risks have practical routes to report them without risking professional or financial consequences.[79] Whistleblowing protections are linked to specified categories of wrongdoing and prescribed reporting routes, such as the Public Interest Disclosure Act 1998.[80] Employees who raise concerns about AI risks may not have a clear route for confidential reporting to a specialist authority.

60. **CONCLUSION**

Employees working on frontier AI systems may be among the first to identify serious safety concerns, but concerns are often raised publicly only after individuals have left their organisations. Contractual restrictions, financial disincentives and uncertainty about appropriate reporting routes discourage employees from raising concerns. Effective AI governance therefore requires clear and protected channels for reporting serious prospective AI risks.

61. **RECOMMENDATION**

The Government should establish protected confidential reporting to the AI regulator. This means reviewing whistleblowing categories, prescribed recipients, worker coverage and remedies for serious prospective AI risks. Lawful reporting must be protected notwithstanding confidentiality clauses.

76 Q1–9
77 Q3
78 Q3
79 Q3; Q5
80 Public Interest Disclosure Act 1998; GOV.uk, Whistleblowing for employees: What is a whistleblower, (Accessed 28 October 2026); GOV.uk, Whistleblowing: list of prescribed people and bodies, (Accessed 28 October 2026);

4 Prevention of harm approach

62. The Online Safety Act 2023 provides an important precedent to inform legislation that reduces digital harms. The Act, for example, requires in-scope service providers to prevent children from accessing certain harmful content and to minimise the risk of vulnerable age groups from encountering other harmful content.[81]

63. The Act was presaged by the Carnegie UK Trust's Internet Harm Reduction proposal. This drew on well-established health and safety legislation and the "precautionary principle" that has informed policy frameworks in other novel areas.[82] One specific precursor considered in the Carnegie paper was the Health and Safety at Work Act 1974, which deals with the employer-employee relationship but which also considers wider harms:

> It shall be the duty of every employer to conduct his undertaking in such a way as to ensure, so far as is reasonably practicable, that persons not in his employment who may be affected thereby are not thereby exposed to risks to their health or safety.[83]

64. Carnegie noted that the concept of a person who "may be affected" presents "a very broad category which, were it to be applied analogously to the online context, could fill the gap in protection". They acknowledge that the 1974 Act "provides a lesser protection to third parties than it does to employees" but argue that the concept of wider harm "serves as a broad model".[84] The Carnegie proposal also notes the challenges in defining "harm" for the purpose of a statutory duty. They note that criminal law, while useful, does not provide a "complete list of harms against which we would expect a service provider to take action".[85] They proposed that Parliament set out "primary or heads of harms as a non-exclusive list to give the regulator initial direction".[86]

81 House of Commons Library, The Online Safety Act: An overview, 30 October 2026
82 Carnegie UK, Internet Harm Reduction, January 2019, para 1
83 Health and Safety at Work etc. Act 1974, section 3(1)
84 Carnegie UK, Internet Harm Reduction, January 2019, paras 33–34
85 Carnegie UK, Internet Harm Reduction, January 2019, para 36
86 Carnegie UK, Internet Harm Reduction, January 2019, paras 47

65. Digital products have become essential to modern life, but unlike almost every other consumer product, they are not governed by the established framework of Product Safety.[87] Recent arguments have advanced a "product safety" approach to novel digital technologies. Tom Collins MP's Digital Product Safety paper, for example, proposes extending a product safety approach to digital products encompassing risk assessment, technical standards, independent assurance and clear legal accountability for providers.

66. Mr Collins argues that a product safety approach would "shift the burden of responsibility from regulators to providers". It envisages a "benchmark expectation of safety" that sits far beneath the threshold of "widespread and serious" harm.[88] The AI Minister, Kanishka Narayan, responded positively to Mr Collins' proposal in the House of Commons Chamber.[89]

67. **CONCLUSION**

A duty to prevent harm and a requirement for safety by design may provide a viable approach for managing AI risks, alongside other measures. Such an approach would require adjustments to tackle the unique risks of AI, but the precautionary approach has long been a feature of safety legislation and recently has informed legislation for online safety.

68. **RECOMMENDATION**

Using the AI Safety and Security Bill, the Government should introduce a duty to prevent harm on developers of AI models. This should cover development, testing, internal deployment, supply, modification and operation, protecting third parties and customers. Duties could follow control over tools, permissions, connections and monitoring. The Bill should include delegated powers to extend into these areas through secondary legislation, following consultation, if required. Protection should continue following a model's release. To safeguard the risks of models developed beyond the UK's jurisdiction, the Government should seek to enshrine this approach in an international convention.

87 Tom Collins MP et al, Digital Product Safety: Proposed Regulation and Standards for a Product Safety approach to Digital, July 2026, p. 2

88 Tom Collins MP et al, Digital Product Safety: Proposed Regulation and Standards for a Product Safety approach to Digital, July 2026, p. 3

89 HC Deb 8 June 2026, c36–37

5 Digital rights

69. The role of a rights-based framework for AI has received increased public and parliamentary attention in recent months. The Joint Committee on Human Rights (JCHR) September 2026 report on 'human rights and the regulation of AI' found that AI presents serious human rights risks, and that these harms may happen rapidly and at scale.[90]

70. The capacity for AI to be used to abuse rights has become increasingly clear. To cite one example, the UN Entity for Gender Equality and the Empowerment of Women has called the proliferation of sexualised deepfake images a "global crisis".[91] Despite recent legislative changes, it was reported in September 2026 that attempts by police to investigate SpaceXAI over sexualised deepfakes generated by its chatbot Grok have been abandoned, and some women targeted by deepfakes are turning to civil litigation to secure accountability.[92]

71. AI also creates risks to individual rights beyond generation of harmful content. Research into human influence indicates that AI systems can change people's expressed attitudes and influence their behaviour. These capabilities also have implications for the effectiveness of human oversight. For example, safety arrangements may depend on people recognising misleading outputs. Evaluation should therefore examine whether AI systems can mislead their overseers or induce users to bypass safeguards.

72. The JCHR noted the lack of coherent AI regulation leaves gaps in which victims have no route to redress, and that there is "too much freedom" for powerful developers to pass on liability. It concluded there is a need for dedicated primary legislation and a new regulator.[93]

90 Joint Committee on Human Rights, Human Rights and the Regulation of AI, HC 160 / HL Paper 56, 14 September 2026, summary

91 UN Women, When justice fails: Why women can't get protection from AI deepfake abuse, 26 February 2026

92 The Observer, The women taking on the tech giants over deepfakes and digital abuse, 29 September 2026

93 Joint Committee on Human Rights, Human Rights and the Regulation of AI, HC 160 / HL Paper 56, 14 September 2026, summary

73. Work on a rights-based AI framework has been underway for several years. The September 2021 Club de Madrid[94] and Boston Global Forum[95] policy lab report on 'fundamental rights in AI and digital societies' provides one example. It called for international rules for AI development that respect human rights and the rule of law.[96]

74. The UK has signed the Council of Europe Framework Convention on AI and Human Rights, Democracy and the Rule of Law. Parties must adopt or maintain measures including respect for dignity and autonomy, transparency, accountability, non-discrimination and privacy.[97] No timetable has been set out for the Convention's ratification and implementation in the UK. JCHR called on the Government to set out a plan for public consultation and ratification.[98]

75. **CONCLUSION**
AI models are increasingly able to independently undertake tasks and perpetrate harms against the individual. Action to protect citizens from AI-facilitated harms must be guided by a focus on established individual rights.

76. **RECOMMENDATION**
The Government should affirm that enshrined individual rights will inform all aspects of government policy on AI regulation. The Government should explain how established rights will be given practical effect through its AI regulatory framework, which should include: proportionate safeguards against deceptive or manipulative uses; requirements for transparency where people interact with AI; and accessible routes to complaint, investigation and redress.

94 Club de Madrid is an international non-profit organisation comprising more than 130 former democratic presidents and prime ministers from over 70 countries who work on issues relating to democracy, governance and global policy challenges.

95 The Boston Global Forum (BGF) offers a venue for leaders, strategists, thinkers, and innovators to contribute to the process of Remaking the World – Toward an Age of Global Enlightenment.

96 Club de Madrid and Boston Global Forum, Policy Lab: Fundamental rights in AI & digital societies: towards an international accord, 7–9 September 2021: final report, pp 10–11

97 Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, Articles 7–13

98 Joint Committee on Human Rights, Human Rights and the Regulation of AI, HC 160 / HL Paper 56, 14 September 2026, para 27

6 The UK's international leadership: a window of opportunity

77. No single nation can keep frontier AI safe.[99] AI firms operate beyond national jurisdictions, and models are difficult for any single country to control.[100] Effective governance will require international cooperation on testing, evaluation, incident reporting and safety standards.[101]

78. While the UK cannot unilaterally regulate overseas frontier AI developers, it has an important role as a convenor, standard-setter and source of technical expertise.[102] The Government highlighted the UK's forthcoming G20 presidency as an opportunity to advance international AI governance.[103] We heard the AISI has given the UK influence in international safety discussions, and the UK could advance common standards and coordination internationally through the growing network of AI safety institutes, which AISI chairs.[104]

79. International cooperation may be slow, incomplete or difficult to enforce.[105] The UK's ability to shape international approaches will depend in part on the credibility of its own arrangements. International leadership should complement, rather than delay, domestic action to address frontier AI risks.

80. **CONCLUSION**
The UK cannot regulate frontier AI developers unilaterally, but it is well placed to shape international approaches to AI safety through technical leadership, standard-setting and international coordination. The pursuit of international agreement should not become an excuse to defer domestic action. The UK's ability to influence international approaches will be strengthened by a coherent domestic framework.

99 Q76
100 Q174
101 Q94; Q96; Q97
102 Q35; Q174
103 Q157; Q171; Q186-Q187
104 Q14; Q90; Q96; Q132; Q143
105 Q35; Q174

81. **RECOMMENDATION**

Within three months, the Government should publish objectives for its G20 presidency and work through the international network of AI safety institutes. These should include common approaches to assessing capabilities, safeguards and residual risks; criteria for escalating scrutiny and intervention; evaluation and monitoring; incident reporting; safety standards; and high-risk open-weight releases. The Government should identify negotiating priorities, responsible ministers and milestones, and report progress alongside domestic implementation. The Government should also assess options for more formal international governance arrangements, including whether there is a case for pursuing a multilateral convention or treaty on advanced AI systems and associated risks. The Committee will examine these objectives through its further work on international AI safety and security.

Conclusions and recommendations

The prize and the peril

1. The Government must address growing safety and security risks without undermining its ability to access frontier systems, lead international standards or benefit from international advances in AI. It must also recognise the limits of domestic regulation in a global ecosystem. (Conclusion, Paragraph 6)

2. Notwithstanding these constraints, the Government must bring forward a comprehensive framework to address the ten critical AI safety and governance gaps we identify in chapter three. These gaps leave significant weaknesses in the UK's safety and security regime and must be addressed, and iterated, in consultation with stakeholders and the public, to build trust in AI and reap the potential economic rewards. (Conclusion, Paragraph 7)

3. The Government must set out a legislative roadmap for AI safety. This should include an AI Safety and Security Bill to enshrine aspects of the new protections we set out in this report. Inevitably, this Bill will need to dovetail with, and enshrine key elements of, an international convention on AI safety which should be the centrepiece of the UK's G20 Presidency. Some elements of the legislative strategy, such as the liability regime, must wait for the Law Commission's findings. A bill in the next session should be merely the first step on that pathway, and the Government should set out the sequence it envisages in response to this report. Given the speed with which this technology is evolving, we accept it may prove essential that ministers are granted order-making powers, which allow them to act to safeguard the public quickly. The transformational nature of the AI revolution will demand that both Government and Parliament are creative when developing new legislative approaches. Parliament should anticipate the creation of new executive powers by designing upfront a strong transparency and accountability framework for granting and holding to account the exercise of these powers. (Recommendation, Paragraph 8)

4. In response to this report, the Government must set out how it will address the disparity in pay between the public and private sector, to recruit and retain people with the technical skills that the public interest requires. (Recommendation, Paragraph 9)

Five approaches to regulating digital risks

5. Democratically elected policymakers should determine AI development and deployment rules. The absence of a clear regulatory framework leaves UK citizens exposed to AI risks, while creating uncertainty for developers and businesses. We find that targeted legislative action, a duty to prevent harm, "digital rights," criminalising the misuse of models for harm, and an international convention may all be needed to manage AI risks and shape opportunities for the future. (Conclusion, Paragraph 13)

6. The Government should set out a coherent legislative strategy for AI safety. This should detail how the different, complementary strengths of the five approaches we outline will be used to develop a coherent regulatory framework. The response to this report should explain how existing legislation, previous legislative commitments, legislation in progress and potential future legislation will tackle the ten risks we identify, alongside efforts to build an international framework. Any proposal must include dedicated new AI legislation to address the challenges we have identified as outside the scope of current or planned provisions. (Recommendation, Paragraph 14)

The gaps

7. Decisions about the risks society should accept are matters for Government, accountable to Parliament, and informed by experts and public consultation. AISI should provide technical evidence to inform those decisions but should not be responsible for determining acceptable risk. A domestic framework must identify who sets risk criteria, who updates these criteria as capabilities and safeguards change, and who acts when those criteria are breached. It must also allow sufficient flexibility to respond to emerging risks. AI firms must be held accountable for operating within these thresholds, and need clear requirements and consequences from the regulator for breaching them. An international convention should seek to agree this approach internationally. (Conclusion, Paragraph 20)

8. The Government should as a first step introduce an AI Safety and Security Bill creating clear legal duties on AI firms to identify and prevent harms from specified high-risk development activities conducted in the UK and specified frontier AI models distributed in the UK. The framework should not

become an unnecessary drag on innovation but should nevertheless assign responsibility for setting and regularly updating risk criteria, informed by AISI's technical assessments and subject to parliamentary scrutiny. It should distinguish reasonable capability thresholds that trigger additional assessment from findings of unacceptable residual risk that require restrictions or intervention. It should specify the safeguards, reporting obligations and regulatory action required at each stage, with a clear jurisdictional basis and proportionate obligations on overseas providers. (Recommendation, Paragraph 21)

9. The Government must set out clear ministerial, and cross Government, accountability for specifying, agreeing and amending harm thresholds, in a way that is transparent and accountable to Parliament. Firms, the public and Parliament are entitled to know who decides what level of risk is tolerable, on what evidence, and with what right of challenge. (Recommendation, Paragraph 22)

10. To advance the safety of frontier models developed abroad and beyond UK statutory control, the Government should seek to enshrine this approach in an international convention on AI safety. (Recommendation, Paragraph 23)

11. The UK should establish enforceable access requirements within its jurisdiction, but recognise domestic legislation will not secure access to models developed overseas. A durable international framework must be introduced alongside our proposed AI Safety and Security Bill, to combine cooperation by overseas developers and governments while preserving the research relationships on which effective evaluation depends. A core objective for the UK's G20 presidency, therefore, should be an international agreement that helps enshrine model access for a trusted network of AI safety institutes and an international consensus around shared approaches to AI assurance and risk management. The international community has developed similar arrangements in, for example, the inspection regime for nuclear energy. We now need similar durable arrangements for AI safety. (Recommendation, Paragraph 27)

12. The Government should continue to resource AISI to enable it to continue to grow its capabilities for frontier analysis. The Government should establish a statutory AI regulator, separate from AISI but drawing on its expertise, with secure multi-year funding. Before legislating, it should set out the regulator's remit, its relationship with AISI, how its powers would be enforced and their likely consequences, including unintended effects. These powers should include access to models when defined thresholds are met or under agreed international conventions; access to relevant systems, logs, training outputs and safeguards; sanctions for unreasonable delays or refusal of access; and ongoing monitoring and re-evaluation. (Recommendation, Paragraph 28)

13. Independent evaluation of frontier AI models is likely to bring significant long-term safety benefits. Any regime for evaluating safety claims will need to provide evaluators with sufficient access and authority to scrutinise both frontier AI systems and the organisational processes through which they are developed, tested, deployed and monitored, throughout their lifecycle. (Conclusion, Paragraph 31)

14. The Government should require developers of designated high-risk systems to substantiate their safety claims and provide independent evaluators with the access necessary to scrutinise them. The framework should establish requirements for evaluator competence, independence, conflicts of interest, reporting and continuing oversight. An international convention should seek to establish a similar regime for model developers beyond the UK's jurisdiction. This framework should support appropriate visibility into model reasoning and chains of thought. (Recommendation, Paragraph 32)

15. Containment failures and other unexpected behaviours can occur during training, testing, evaluation and internal deployment, before any public release decision is taken. Controls focused solely on externally deployed systems leave a major gap. Future governance arrangements should provide appropriate visibility and oversight of high-risk internal deployment and evaluation activities. (Conclusion, Paragraph 34)

16. Through the AI Safety and Security Bill, the Government should legislate to ensure that its domestic regulatory framework covers high-risk training, testing, evaluation and internal deployment activities, as well as externally deployed systems. Working with the AI Security Institute, regulators and industry, it should develop guidance and, where appropriate, minimum standards for the containment, monitoring and oversight of high-risk frontier AI systems. Any new AI regulator should be endowed with powers to obtain information about containment arrangements, testing activities, incidents and mitigations associated with deployment and evaluation. To guard against such risks in models developed abroad, the UK should seek to enshrine this approach in an international convention. (Recommendation, Paragraph 35)

17. AI firms should not be exclusively responsible for determining whether their frontier systems are safe to deploy. The Government must therefore create a risk and regulation framework that both optimises opportunity and manages risk. While firms should have to evidence their safety claims, designated high-risk systems should require independent safety verification, informed by evaluation throughout development, testing and internal use. Evaluators must have access to test developers' claims and the ability to disclose where access has been restricted or their findings differ. (Conclusion, Paragraph 38)

18. The Government should explore how an independently scrutinised safety assessment could assess designated high-risk systems before they are deployed in the UK. This framework should include an assessment of the system's capabilities, the effectiveness of safeguards and the risks remaining in its intended circumstances of use. It should demonstrate compliance with applicable risk criteria, record material disagreements between developers and evaluators, and specify conditions for deployment. Material changes in capability, safeguards, controllability or use should trigger reassessment, with restrictions or suspension where risks cannot be adequately controlled. (Recommendation, Paragraph 39)

19. The Government cannot assess frontier AI risk without systematic information about failures and near misses. Fragmented voluntary arrangements prevent collective learning and may leave individual companies to decide unilaterally whether an event is sufficiently serious to disclose. (Conclusion, Paragraph 44)

20. Using the AI Safety and Security Bill, the Government must introduce mandatory reporting of frontier AI incidents of material concern to the UK. It should seek to internationalise these arrangements through an international convention. This approach should be supported by a clear and standardised reporting framework. This framework could include:

- clearly defined reportable incidents and near misses;
- standardised severity classifications;
- clearly defined notification timelines;
- clear requirements on who must report incidents and to which authority;
- requirements to report the extent of visibility into model behaviour and reasoning processes;
- legal protections for confidential and security-sensitive reporting; and
- requirements for corrective action, follow-up reporting and resolution updates. (Recommendation, Paragraph 45)

21. No UK regulator or public authority currently has the power to impose safeguards on frontier AI systems, pause training, delay deployment or withdraw unsafe services. Decisions about whether systems are safe to develop or deploy remain with the companies building them. (Conclusion, Paragraph 48)

22. Where risk cannot be adequately controlled, the Government should ensure that the AI safety regulator defines and seeks powers to require safeguards for AI products serving UK users. In defined circumstances where the Government has significant concerns about an AI model, it should have the power to delay, restrict or prohibit its deployment, training or testing in the UK. The statutory regulator should not be AISI. (Recommendation, Paragraph 49)

23. Safety frameworks cannot assume that all material risks will be identified before deployment. Once model weights have been widely copied and distributed, recalling every copy or preventing all further use may be technically impossible. However, it is important that Government has the capability to target action at particular hosts, repositories, services or unlawful uses, and to criminalise the harmful misuse of models, or acts that enable such misuse. In addition, to control misuse of irretrievable models, the Government may have to develop a way to criminalise the misuse of models for harm. (Conclusion, Paragraph 52)

24. Given the difficulty of reversing distribution, where it has jurisdiction, the Government should require independent scrutiny before high-risk open-weight models are released. AI firms and relevant hosts and repositories should be required to maintain and exercise proportionate emergency plans with regulators, operational agencies and international partners. These plans should distinguish between systems remaining under an identifiable operator's control (where access can be restricted or services suspended), and distributed models that cannot be recalled. The Government must have in place capability to detect and remove harmful models from services where it can, to disrupt unlawful use, and ensure the right range of criminal offences are available to deter the financing of, advertisement of, distribution of, or misuse of AI models. This Committee will explore the best means to develop and deliver an international AI framework that supports these powers in our future work. (Recommendation, Paragraph 53)

25. A safety regime is incomplete if responsibility remains unclear when safeguards fail or harm occurs, and those harmed do not have clear and effective routes to redress. Clarifying liability arrangements, and ensuring AI firms cooperate with them, will require international collaboration, common standards and effective enforcement. (Conclusion, Paragraph 56)

26. The Law Commission will report on its product liability consultation in due course. Following the Law Commission's report, the Government should work with international partners to clarify the responsibilities of developers, deployers, infrastructure providers and users where AI systems cause harm or safeguards fail. In addition, the Government should identify outstanding matters and gaps outside the scope of the Law Commission's consultation and set out how it intends to address those

gaps. Ministers should coordinate with devolved administrations and set a timetable for responding to the Law Commission's recommendations. (Recommendation, Paragraph 57)

27. Employees working on frontier AI systems may be among the first to identify serious safety concerns, but concerns are often raised publicly only after individuals have left their organisations. Contractual restrictions, financial disincentives and uncertainty about appropriate reporting routes discourage employees from raising concerns. Effective AI governance therefore requires clear and protected channels for reporting serious prospective AI risks. (Conclusion, Paragraph 60)

28. The Government should establish protected confidential reporting to the AI regulator. This means reviewing whistleblowing categories, prescribed recipients, worker coverage and remedies for serious prospective AI risks. Lawful reporting must be protected notwithstanding confidentiality clauses. (Recommendation, Paragraph 61)

Prevention of harm approach

29. A duty to prevent harm and a requirement for safety by design may provide a viable approach for managing AI risks, alongside other measures. Such an approach would require adjustments to tackle the unique risks of AI, but the precautionary approach has long been a feature of safety legislation and recently has informed legislation for online safety. (Conclusion, Paragraph 67)

30. Using the AI Safety and Security Bill, the Government should introduce a duty to prevent harm on developers of AI models. This should cover development, testing, internal deployment, supply, modification and operation, protecting third parties and customers. Duties could follow control over tools, permissions, connections and monitoring. The Bill should include delegated powers to extend into these areas through secondary legislation, following consultation, if required. Protection should continue following a model's release. To safeguard the risks of models developed beyond the UK's jurisdiction, the Government should seek to enshrine this approach in an international convention. (Recommendation, Paragraph 68)

Digital rights

31. AI models are increasingly able to independently undertake tasks and perpetrate harms against the individual. Action to protect citizens from AI-facilitated harms must be guided by a focus on established individual rights. (Conclusion, Paragraph 75)

32. The Government should affirm that enshrined individual rights will inform all aspects of government policy on AI regulation. The Government should explain how established rights will be given practical effect through its AI regulatory framework, which should include: proportionate safeguards against deceptive or manipulative uses; requirements for transparency where people interact with AI; and accessible routes to complaint, investigation and redress. (Recommendation, Paragraph 76)

The UK's international leadership: a window of opportunity

33. The UK cannot regulate frontier AI developers unilaterally, but it is well placed to shape international approaches to AI safety through technical leadership, standard-setting and international coordination. The pursuit of international agreement should not become an excuse to defer domestic action. The UK's ability to influence international approaches will be strengthened by a coherent domestic framework. (Conclusion, Paragraph 80)

34. Within three months, the Government should publish objectives for its G20 presidency and work through the international network of AI safety institutes. These should include common approaches to assessing capabilities, safeguards and residual risks; criteria for escalating scrutiny and intervention; evaluation and monitoring; incident reporting; safety standards; and high-risk open-weight releases. The Government should identify negotiating priorities, responsible ministers and milestones, and report progress alongside domestic implementation. The Government should also assess options for more formal international governance arrangements, including whether there is a case for pursuing a multilateral convention or treaty on advanced AI systems and associated risks. The Committee will examine these objectives through its further work on international AI safety and security. (Recommendation, Paragraph 81)

Formal minutes

Tuesday 10 November 2026

Members present:

Liam Byrne, in the Chair

Dan Aldridge

Antonia Bance

Chris Bloore

Nesil Caliskan

Leigh Ingham

Charlie Maynard

Mr Joshua Reynolds

Guardrails for growth: AI safety and security

Draft Report (*Guardrails for growth: AI safety and security*), proposed by the Chair, brought up and read.

Ordered, That the draft Report be read a second time, paragraph by paragraph.

Paragraphs 1 to 81, read and agreed to.

Summary agreed to.

Resolved, That the Report be the Eleventh Report of the Committee to the House.

Ordered, That the Chair make the Report to the House.

Ordered, That embargoed copies of the Report be made available, in accordance with the provisions of Standing Order No. 134.

Adjournment

Adjourned till Wednesday 11 November at 1.45pm

Witnesses

The following witnesses gave evidence. Transcripts can be viewed on the inquiry publications page of the Committee's website.

Tuesday 13 October 2026

Daniel Kokotajlo, Executive Director, AI Futures Project; **Aengus Lynch**, Technical Staff, Theorem; **Steven Adler**, Co-Founder and Chief Scientist, Guidelight AI Standards; **Geoffrey Irving**, Co-Founder and Chief Scientist, Resolution Q1–35

The Lord Booth-Smith, Head of Europe, Middle East and Africa Policy, Anthropic; **Tom Duff Gordon**, Vice-President and Head of Europe, Middle East and Africa Policy, OpenAI; **Owen Larter**, Senior Director, Google DeepMind; **Rob Sherman**, Vice President, Global Policy, Meta Q36–152

Henry De Zoete, Director, AI Security Institute; **Cat Little CB**, Permanent Secretary, Cabinet Office; **Alexandra Jones**, Director General Growth, International, Science & Technology, Department for Business, Innovation, Science and Trade Q153–188

Published written evidence

The following written evidence was received and can be viewed on the inquiry publications page of the Committee's website.

AIS numbers are generated by the evidence processing system and so may not be complete.

1	Anthropic	AIS0004
2	ControlAI	AIS0001
3	Meta Platforms, Inc.	AIS0002
4	OpenAI	AIS0003

List of Reports from the Committee during the current Parliament

All publications from the Committee are available on the publications page of the Committee's website.

Session 2026–27

Number	Title	Reference
10th	China and the UK economy: the case for risk-managed growth	HC 124
9th	Critical minerals	HC 132
8th	Pre-appointment hearing with the Government's preferred candidate for the Chair of the Financial Reporting Council	HC 540
7th	Small business strategy: Government Response	HC 122
6th	The UK-US economic relationship: 250 years on	HC 127
5th	Post Office Horizon scandal: justice for sub-postmasters: Government Response	HC 129
4th	UK-EU relations a decade on	HC 126
3rd	Competition and market functioning in the UK live music industry	HC 128
2nd	The regulation of postal services	HC 131
1st	Investing in the UK economy	HC 123
5th Special	Small business strategy: Second Government Response	HC 748
4th Special	The UK-US economic relationship: 250 years on: Government Response	HC 672
3rd Special	Competition and market functioning in the UK live music industry: Government Response	HC 625
2nd Special	The regulation of postal services: Government Response	HC 624

Number	Title	Reference
1st Special	Investing in the UK economy: Government Response	HC 623

Session 2024–26

Number	Title	Reference
17th	Pre-appointment hearing with the Government's preferred candidate for the Chair of the Competition and Markets Authority	HC 1723
16th	Post Office Horizon scandal: Justice for sub-postmasters	HC 1589
15th	Small business strategy	HC 1057
14th	Toward a new doctrine for economic security: Government Response	HC 1666
13th	Priorities of the Business and Trade Committee for 2026	HC 1411
12th	UK-India Comprehensive Economic and Trade Agreement (CETA)	HC 996
11th	Toward a new doctrine for economic security	HC 835
10th	US Economic Prosperity Deal	HC 1306
9th	Draft Legislative Reform (Disclosure of Adult Social Care Data) Order 2025	HC 1140
8th	Export led growth: Trade with the Asia-Pacific region	HC 1048
7th	Industrial Strategy	HC 727
6th	How to strengthen UK-EU relations: Policy Priorities for the Summit	HC 908
5th	How to strengthen UK-EU relations	HC – 814
4th	Post Office Horizon scandal redress: Unfinished business: Government response	HC 778
3rd	Make Work Pay: Employment Rights Bill	HC 370
2nd	Priorities of the Business and Trade Committee	HC 423
1st	Post Office and Horizon scandal redress: Unfinished business	HC 341
7th Special	UK-India Comprehensive Economic and Trade Agreement (CETA): Government Response	HC 1832
6th Special	US Economic Prosperity Deal: Government Response	HC 1545
5th Special	Export led growth: Trade with the Asia-Pacific Region: Government Response	HC 1324

Number	Title	Reference
4th Special	Industrial Strategy: Government Response	HC 1305
3rd Special	How to strengthen UK-EU relations: Policy Priorities for the Summit: Government Response	HC 1267
2nd Special	Post Office Horizon scandal redress: Unfinished business: Government response	HC 969
1st Special	Make Work Pay: Employment Rights Bill: Government response	HC 932