AI Developer Liability for Severe Harms

Policy Briefing

AI poses substantial risks to national security

In April 2026, the AI company Anthropic announced a new model called Claude Mythos that is substantially more capable than any previous model at cyber security tasks. Research by the UK AI Security Institute found that it can "discover and exploit vulnerabilities autonomously – tasks that would take human professionals days of work." Mozilla recently announced that in the past month they had "identified and fixed an unprecedented number of latent security bugs in Firefox with the help of Claude Mythos Preview and other AI models."

In February 2026, researchers at Gambit Security discovered that "a single operator used AI to breach nine Mexican government organisations and exfiltrate hundreds of millions of citizen records." Their report found that "approximately 75 percent of remote command execution activity across the campaign was generated and executed by Claude Code." What was once the domain of highly sophisticated state-sponsored hacking groups is now achievable by a single motivated individual with access to AI models.

A similar attack directed at UK critical infrastructure could compromise services on which millions of British citizens depend, with consequences for public safety, financial stability and national security. According to the Department for Science, Innovation and Technology, "cyber attacks can lead to unsafe drinking water, no electricity, hospitals unable to access digital patient records, and businesses unable to access their systems." In 2017, the NHS was a victim of the WannaCry attack, which disrupted 34% of trusts in England and caused the cancellation of 19,000 appointments.

AI models continue to improve at an exponential rate with no end in sight. The coming months and years will bring new models which are much more capable again than Claude

Mythos. The UK must catch up to this growing threat before the public suffers the consequences.

Proposed cyber security legislation is not sufficient

The government has recognised the seriousness of cyber attacks and introduced the Cyber Security and Resilience Bill. This is a welcome update to existing regulations and will strengthen UK essential services. But the Bill only regulates the targets of attack, not the producers of offensive AI capabilities.

The compliance burden is placed on water utilities, NHS trusts and other operators to improve their cyber defences. The standard of "appropriate and proportionate" security, which the Bill requires, is now a moving target. As explained above, AI capabilities are transforming the cyber landscape over the span of mere months. UK essential services will not be able to keep pace with the breakneck speed at which frontier AI is evolving. The government must shift some of the responsibility back to the AI companies that are increasing cyber risk.

Beyond cyber threats

Today's frontier AI is a general-purpose technology, and other types of severe risks are likely to emerge in the near future. The International AI Safety Report 2026 found that existing AI systems can offer "expert-level laboratory instructions" for creating biological and chemical weapons. OpenAI's research found that its own models are "on the cusp of being able to meaningfully help novices create known biological threats."

Many AI experts have warned of the potential for humans to lose control of AI systems and the catastrophic risks that superhuman AI could pose. In 2023, hundreds of leading figures in AI signed a joint statement declaring that "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."

UK legislation should not try to forecast which threats are most likely to emerge. It should instead align the incentives of AI developers to detect and mitigate all severe harms so that the public will be protected, whatever path AI development may take.

AI developers are not currently accountable for the risks they create

Under the status quo, there is substantial doubt as to whether model developers have any liability for even the most severe and foreseeable harms, including cases in which their models were clearly decisive in enabling the harm. The Law Commission has warned that autonomous AI systems could create "liability gaps" where no one is liable for harm caused by an AI system.

Causation will be difficult to establish under existing liability frameworks such as the Consumer Protection Act 1987 or the Online Safety Act 2023. Nicola Cain, CEO of legal consultancy Handley Gill, commenting on the UK Jurisdiction Taskforce's draft Legal Statement on Liability for AI Harms, notes "a dichotomy whereby the less responsibly developers of AI systems [...] behave, the less likely that liability will attach." This is because causation is harder to establish when models are deployed without safety testing as developers can claim that emergent behaviour was unforeseeable. We must introduce a new framework that aligns the incentives of developers with the public interest.

Simple legislation for a complex technology

AI capabilities are advancing at an extraordinary pace. It is difficult to design strong legislation that keeps up with the evolving risks of the technology and avoids stifling responsible innovation. Liability is a simple mechanism that gives AI developers the incentives to put their full effort into developing AI responsibly with the best methods available.

Where prescriptive regulation must chase evolving technologies, liability establishes a durable principle: those who profit from creating risk must bear meaningful responsibility when that risk materialises. The principle is technology-neutral, future-proof and consistent with the way UK law has long treated other dangerous industries.

There are many potential threats posed by AI systems, ranging from cyber attacks and fraud to bioweapon development and loss of control. Parliament does not need to weigh in on scientific questions around the likelihood of each threat. Liability would be based on the degree of harm, not the mechanism through which harm occurred.

Under this regime, model developers can choose to implement only light-touch precautions against the threats they are confident are minimal. And they will be able to justify substantial investments to mitigate those risks that they judge to be truly severe.

A liability regime also sidesteps a key uncertainty that accompanies many other types of legislation that aim to address the risks of AI. Since jurisdiction is based on where the harm occurs rather than where the developer is located, foreign developers whose models cause harm in the UK would be liable regardless of whether they continued to serve the UK market. As all three of the leading frontier AI developers (OpenAI, Google DeepMind and Anthropic) have substantial assets in London, they could not credibly withdraw to escape the practical reach of UK enforcement. Any attempt to do so would itself be a striking signal that developers consider the underlying risks too severe, and their own ability to mitigate them too limited, to accept the liability.

While the detailed mechanics will benefit from further consultation, the design questions a workable bill must address are well understood. Among them: how causation is established where harm flows through a chain of developer, deployer and user; what threshold of severity should trigger liability; and how the evidential burden should be allocated given the asymmetry of information between developers and those affected by their products. These are tractable questions, well precedented in adjacent areas of British law.

Enforcement raises a second set of questions, particularly given the concentration of frontier AI development outside the United Kingdom. A range of mechanisms are available, and the experience of other regulatory regimes that have extended jurisdiction over foreign companies (both successful and unsuccessful) will inform the choices made. Collective redress mechanisms will also require careful attention, since the most serious AI-enabled harms may be spread across many individuals, or so severe that no single claimant is positioned to litigate effectively. The right combination of design choices will emerge through consultation with legal practitioners, industry and Parliament.