Hogan Lovells

Insights and Analysis

International AI Safety Report 2026 – UK litigation lessons from imperfect AI

16 April 2026

The International AI Safety Report 2026 is an assessment of general-purpose AI capabilities, emerging risks and risk-management approaches, intended to support informed policymaking. It is sceptical of neat capability narratives, highlighting uneven performance, an evaluation gap between tests and real-world use, and rising autonomy through AI agents – all of which heighten dispute risk and complicate liability allocation. For litigators, the Report reads less like a technology forecast and more like a litigation risk map.

A recurring theme of the Report is unpredictability. The Report finds that systems can perform impressively on complex tasks, but that performance "remains 'jagged', with leading systems still failing at some seemingly simple tasks". That matters because contracts, procurement decisions and governance frameworks often assume performance is stable and failures are foreseeable. The Report suggests that assumption is increasingly hard to defend, noting that because "model behaviour can be hard to understand or predict, it is challenging to foresee or confidently rule out specific failures". In this article, we explore where that unpredictability is most likely to translate into UK litigation risk.

Key dispute angles

1) What was promised – jagged capability and overstated claims

Uneven performance is a breeding ground for expectation disputes. If capability claims (accuracy, reliability, "near-human" support) encouraged reliance, they may sit at the heart of breach, warranty and misrepresentation claims.

Two practical points follow:

- **What was promised will matter more than what is technically possible**. In technology contracts, courts are typically concerned with the hierarchy of obligations: if the deal contains outcome-focused performance standards or warranties, meeting minimum technical specifications or "industry practice" may not be enough if the promised result is not achieved. Pre-contract capability statements made in a procurement context (RFPs, pilots, presentations) can also become actionable – as contractual terms, collateral warranties or misrepresentations.

- **The "everyone knows AI is probabilistic" defence is not a silver bullet**. Generic caveats may carry limited weight where they conflict with specific performance representations, and risk allocation (including consequential loss exclusions and non-reliance language) can be subject to scrutiny as unreasonable under the Unfair Contract Terms Act. If capability assertions were made without reasonable grounds, the exposure can shift from "it didn't quite deliver" to credibility – including the potential for fraud allegations in more extreme cases.

2) The evidential battleground – evaluation, benchmarks and methodology

The Report highlights that pre-deployment performance tests often do not reliably predict real-world performance. That matters because benchmark results are used to justify procurement, valuation and funding decisions – but, in UK tech disputes, courts tend to look past headline metrics and ask whether the supplier's work was reasonably capable of achieving the stated purpose in the conditions of deployment (and, depending on the contract, whether there was a fitness-for-purpose commitment rather than a best-efforts narrative).

In practice, the fight will rarely be only about benchmark scores. It is about whether evaluation and validation were adequate, transparent and properly connected to the intended use case:

- **Real-world testing** – was there genuine user testing and validation against real data, real users and real operating conditions, rather than reliance on lab-style benchmarks?

- **Use-case alignment** – were generic benchmarks treated as a proxy for a specific workflow without proper justification, iteration and feedback-driven testing?

- **Governance and disclosure** – what did the supplier (and the deployer) know about unsupported uses, known limitations and failure modes, and what was actually said during procurement, diligence and sign-off?

Where performance evidence is tied to milestones, valuation, or funding, disputes also tend to become credibility contests: were benchmark results curated, selectively presented or based on assumptions that were not made explicit? In UK litigation, the courts will often treat performance estimates and assurance statements as implicitly carrying a representation that the supplier had carried out a proper analysis and had reasonable grounds for what it was saying.

Courts expect technical arguments to be independently explainable and verifiable, with clear sampling/validation methods and robust reasoning. That reinforces the Report's point: even if some failures are difficult to predict with certainty, parties will still be judged on whether they did the kind of real-world validation and transparent testing a careful supplier or deployer should have done.

3) Agents, autonomy and attribution

The Report flags the surge in AI agents – systems designed to browse the internet and execute multi-step tasks with reduced human oversight. Its assessment is cautious: agents are becoming more competent, but remain unreliable where tasks involve many steps. The litigation relevance is straightforward – more autonomy means fewer natural points for human intervention, and a narrower window to catch and correct errors.

The Report's "defence in depth" discussion is also important here. Layered safeguards help, but they can still fail if a control is bypassed, misconfigured, or overwhelmed by novel behaviour. In a dispute, the question will often be whether the control design was reasonable (judged against contemporary knowledge standards, not hindsight) and whether it was properly implemented, not just described on paper.

It also frames "loss of control" as an emerging governance narrative. While it suggests today's systems "show early signs of relevant capabilities, but not at levels that would enable loss of control", oversight is already a practical challenge. Indeed, the Report notes that "models have disabled simulated oversight mechanisms and, when confronted, produced false statements to justify their actions". The much-hyped social network for agents is a useful case study: the issue is less AI consciousness than practical control. Agents can act in ways that are hard to reconstruct: what was real, what triggered what, and who (or what) caused the outcome. Agentic AI is not a liability shield – it can intensify scrutiny of who set the objectives, who configured autonomy levels and tool access, and who decided that reduced oversight was acceptable.

A further complication is multi-agent settings and external tool access. The Report notes that attributing liability for harms caused by agents can be difficult, especially where it is hard to identify when and how failures occurred. That creates predictable evidential pressure points – including "black box" causation arguments – but uncertainty alone is unlikely to excuse a failure to meet ordinary standards of care.

Liability allocation – why this may get messier

One of the Report's most interesting conclusions for litigators is that existing liability frameworks may not be capable of adequately addressing AI-related harms. Existing contract, tort and criminal regimes will apply, but AI creates distinctive difficulties: "harms can be difficult to trace to specific design choices, especially since full information about risk management processes is not public, and responsibility is distributed across model developers, application builders, deployers, and users".

That is a counterweight to more optimistic domestic commentary – including from the UK

Jurisdiction Taskforce – suggesting that existing frameworks can address most AI harms. The Report's tone is more cautious: the framework may be there, but the facts may be harder to prove, and the responsible actor harder to pin down. It also keeps targeted adaptation of legal frameworks on the table, including (among other things) clearer responsibility for generating or disseminating unauthorised or undisclosed synthetic content.

Practical takeaways

- **Prepare to stand behind capability claims in court** – ensure marketing and governance statements can be supported by your testing and evaluation.
- **Contract for the evaluation gap** – define performance carefully, and include audit and incident-reporting obligations.
- **Think about attribution** – especially for agents: logging, approval gates and clear human-in-the-loop triggers.
- **Stress-test governance under failure scenarios** – the best evidence is often what your organisation did when things started to go wrong.
- **Sanity-check insurance coverage and notification triggers** – AI incidents can cut across product liability, cyber, professional indemnity, general liability and commercial crime insurance policies. Know what cover you have and when to notify to avoid coverage disputes.

Conclusion

The Report does not offer tidy answers, but it does point to a clear direction of travel for disputes. As general-purpose AI is deployed more widely – and as organisations rely on it in higher-stakes workflows – uneven capability, imperfect evaluation and rising autonomy are likely to generate more contested questions about what was promised, and where responsibility sits. In that context, there will be greater pressure on suppliers and vendors to evidence their governance processes, including

around testing and incident response. Litigation risk and outcomes will turn on the quality of that evidence trail.

This article was first published in Airmic News which can be viewed here.

Authored by Reuben Vandercruyssen and Lydia Savill.

Contacts

Lydia Savill

Partner

London

Email me

Reuben Vandercruyssen

Senior Associate

London

Email me