Enabling AI Safety Information Sharing: UK Competition Law Block Exemptions and Institutional Design

Anonymous Author(s)
Affiliation
Address
email

Abstract

Frontier AI labs face a coordination failure: sharing safety-critical information could prevent systemic failures, but competition law, designed to prevent collusion, creates legal barriers to collaboration. This paper addresses this coordination failure through comparative institutional analysis and legal framework redesign for the UK. Drawing on cybersecurity (ISACs/ISAOs, CISA) and pharmaceutical (EudraVigilance) precedents from the UK, EU, and US, we demonstrate how sector-specific legal exemptions paired with neutral clearinghouse institutions resolve tensions between competition enforcement and safety-critical information exchanges. We develop a two-dimensional taxonomy that maps technical AI information by commercial sensitivity and safety relevance, enabling clearinghouses and competition authorities to weigh antitrust risk against safety value. Analysing UK Competition Act Chapter I reveals that safety-critical information exchanges currently lack legal clarity; most crucially, the existing R&D Block Exemption Order (2022) does not protect post-deployment disclosures. Our analysis demonstrates that effective block exemptions require three design principles: (1) FRAND access, (2) anonymisation through neutral intermediaries, and (3) transparency requirements. We propose establishing the UK AI Security Institute (AISI) as a neutral clearinghouse and systematically evaluate nine institutional mechanisms to incentivise AI lab information sharing.

1 Introduction

Sharing information about model vulnerabilities and risk mitigation techniques could prevent systemic AI failures, yet UK competition law lacks clarity on when such collaboration is permissible. Labs sharing dangerous capabilities discoveries, red-teaming results, or safety protocols risk violating Chapter I prohibitions of the Competition Act 1998, facing potential fines of up to 10% of worldwide turnover for behaviour that serves public safety. This legal uncertainty persists despite widespread recognition among researchers and industry that catastrophic AI failures could damage the entire sector.

This paper examines how competing developers share safety-critical information about model vulnerabilities, dangerous capabilities, and mitigation techniques without running afoul of competition law. We focus on optimal collaboration in the UK context, where the Competition and Markets Authority (CMA) has signaled increased scrutiny of AI foundation model markets while, at the same time, the UK AI Security Institute (AISI) seeks to position itself as a global leader in AI safety research.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

1.1 Contributions

This paper provides the first comprehensive legal analysis of competition law exemptions specifically designed for AI safety information sharing in the UK context. Our three novel contributions are: (1) a two-dimensional taxonomy mapping AI-specific information by commercial sensitivity and safety relevance across model development stages (Section 3.2), (2) detailed application of UK Competition Act Chapter I analysis to AI safety scenarios including a worked example of chain-of-thought sharing (Section 3.3), and (3) concrete institutional design analysis establishing AISI as neutral clearinghouse with systematic evaluation of nine complementary incentive mechanisms (Section 4). The legal framework analysis and comparative evaluation represent entirely new empirical and analytical work not present in prior publications.

1.2 The information sharing dilemma

Frontier AI developers' coordination problem has the characteristics of both a prisoner's dilemma and a public goods provision challenge. Sharing critical safety information creates positive externalities: all labs benefit from collective knowledge about risks to mitigate harmful tendencies and capabilities of AI (bias, deceptiveness, the ability to support cyberattacks). However, several interrelated factors discourage voluntary disclosure:

Competition law uncertainty. Information exchanges between competitors may trigger antitrust scrutiny under the UK Competition Act 1998 Chapter I prohibitions. The legal boundaries are unclear for AI safety information sharing because: (1) safety research often correlates with competitive advantage, (2) information about model capabilities and timelines can signal market conduct even when framed as safety disclosures, and (3) there is no precedent, and a lack of in-house technical expertise, for how competition authorities should evaluate safety-motivated information sharing in rapidly evolving technology sectors. Labs face significant legal risk including fines reaching up to 10% of turnover (1998).

Competitive disadvantage. A laboratory that invests millions in developing sophisticated evaluation benchmarks or novel alignment techniques and shares these findings enables competitors to incorporate insights at minimal cost, creating first-mover disadvantage.

Liability exposure. Disclosing safety issues creates documentary trails, establishing foreseeability in future tort or criminal proceedings. Under UK law, liability for negligence requires demonstrating that harm was foreseeable and reasonable precautions were not taken. By keeping information internal, labs reduce the likelihood of legal action since potential claimants may lack evidence needed to justify claims.

Misaligned market incentives. Current market structures reward capability advancement over safety investment. Consumer purchasing decisions are driven by visible capabilities rather than safety practices, creating weak market incentives for voluntary safety investments.

To locate solutions that weaken these disincentives, we examine how industries with similar characteristics have historically shared information and attempted private governance. We analyse case studies from cybersecurity for its vast technical information overlap, pharmaceuticals for its societal recognition as a public good, and digital advertising for its twenty-first century emerging-industry profile.

2 Precedents from safety-critical industries

As a sector, AI is new, large, dominated by limited major players, but still capable of sudden, rapid evolution. Contemporary case studies provide insights from past antitrust challenges and the merits of private or quasi-governmental governance structures.

2.1 Cybersecurity: ISACs, ISAOs and CISA

In 2001, nineteen U.S. high-tech companies formed the Information Technology Sharing and Analysis Centre (IT-ISAC) following Presidential Decision Directive 63, which recognized that critical infrastructure security would benefit from sharing cybersecurity risk information (1998). ISACs exchanged threat intelligence in partnership with government agencies (DHS, FBI) but operated as closed networks.

During the 2008-2009 Conficker worm crisis, the IT-ISAC mobilized major IT firms to share real-time threat intelligence via anonymised channels, facilitating the sink-holing of millions of malicious domains (2010). This demonstrated how voluntary, anonymised information sharing under strong institutional governance can unify competing firms during collective risk. The UK currently lacks a directly equivalent framework to CISA.

In 2015, Executive Order 13691 created ISAOs—flexible counterparts enabling any community to form information sharing bodies with lower barriers to entry (2015).

That same year, the Cybersecurity Information Sharing Act (CISA) was enacted, providing three key protection categories: **(1) Antitrust exemptions**—Section 104(E) establishes that sharing information on threats or defences is not considered anti-competitive, and Section 106(B) provides that entities sharing information are protected from antitrust liability; **(2) Liability protections**—companies sharing cyber threat indicators in good faith receive liability shields; **(3) Confidentiality assurances**—Section 105(D)(2) protects information shared with government as proprietary and not subject to FOIA requests (2015). In 2018, CISA agency developed Automated Indicator Sharing (AIS) enabling real-time exchange with anonymisation capabilities (2018).

2.2 Pharmaceuticals: pharmacovigilance

In the EU, the pharmaceutical sector shares safety information through European Medicines Agency (EMA)-facilitated networks (2023). Companies must legally report adverse drug reactions to regulators, who share findings industry-wide through EudraVigilance. This pharmacovigilance system is hybrid: industry generates data while government (EMA's Pharmacovigilance Risk Assessment Committee) facilitates analysis and dissemination (2025). Because reporting is mandatory, companies participate equally. EudraVigilance allows companies to see industry-wide safety trends without exposing which competitor contributed which report. When serious risks are identified, all manufacturers of similar drugs are alerted simultaneously.

Thus, in the EU's pharmaceutical sector, patient safety trumps competitive secrecy: regulations ensure critical risk information flows to all relevant parties. The precedent of regulated information-sharing schemes (with legal mandates or incentives to report "adverse events" and incidents) may be effective in balancing transparency and competition in other sectors.

2.3 Digital advertising: self-regulation limits

In the early 2000s, leading ad networks formed the Network Advertising Initiative (NAI) as a self-regulatory trade association before formal privacy regulations. As voluntary initiative, NAI avoided antitrust pitfalls by limiting collaboration to privacy and consumer protection, and receiving endorsement from the Federal Trade Commission (2000). However, over time stronger U.S. state-level laws emerged and NAI's influence waned. This illustrates how industry-led governance can buy regulatory goodwill and shape early norms, but decentralization can shift a trade association's external validation and legal reinforcement to remain credible.

As of March 2025, the Frontier Model Forum (FMF), a trade association that also sets pre-regulatory safety standards, exists in a juxtaposed capacity to the NAI for digital advertising. Except that the FMF is also only for members, who are presently only frontier AI firms. The NAI case study suggests the FMF is valuable as governments catch up in learning how a new industry fits into the present economy, but self-regulation becomes symbolic over time.

3 Competition law analysis

3.1 UK legal framework

The CMA possesses significant powers under the Competition Act 1998 to investigate and penalise anti-competitive agreements falling into Chapter I prohibitions. The CMA's April 2024 market analysis focuses on AI foundation models, explicitly identifying risks of incumbent firms restricting access to critical inputs. Existing frameworks like the R&D Block Exemption Order 2022 offer potential routes for permissible collaboration, but applicability to AI safety information sharing is unclear. The primary barrier is the current absence of specific CMA guidance that explicitly clarifies which types of AI safety information sharing are considered low-risk.

3.2 Information classification framework

We develop a two-dimensional framework to classify frontier AI lab information by commercial sensitivity (CSI) and safety relevance across the pre-training/during training and post-deployment phases.

Table 1: Risk matrix: before/during training

	Low CSI	**High CSI**
Low Safety	Irrelevant	Proprietary datasets, Compute ownership
High Safety	Safety evaluation plans, Red team structures	Compute usage, Risk thresholds, Capabilities

Table 2: Risk matrix: after deployment

	Low CSI	**High CSI**
Low Safety	Irrelevant	Deployment timelines, Monetization
High Safety	Refusal accuracy, Jailbreak prevention	Red-teaming discoveries, Vulnerabilities

Information shared after deployment is generally less likely to qualify under R&D exemption and must be assessed with greater legal scrutiny.

3.3 Technical challenges of generating and sharing types of information

The technical challenges of generating safety-relevant information vary substantially across our classification framework (Tables 1-2). For high-safety, high-CSI information such as red-teaming discoveries or novel vulnerability identification, the primary challenge lies in capability elicitation. Ensuring that safety evaluations truly uncover the model's dangerous capabilities, rather than merely testing what developers already know to look for, and preventing sandbagging of increasingly situationally aware models. Additionally, generating meaningful safety information about emergent capabilities requires extensive computing resources for comprehensive testing across diverse scenarios, specialised expertise in both AI systems and specific risk domains (CBRN, cybersecurity, autonomous weapons), and sophisticated instrumentation to detect subtle behavioural patterns that might indicate deceptive alignment or hidden capabilities. The generation challenge is compounded by the fact that as models become more capable, the search space for potential failure modes expands, and the time required for thorough safety evaluation may conflict with commercial pressures for rapid deployment. Furthermore, some types of information require technological innovation, such as jailbreaking prevention, redteaming structures or refusal accuracy.

Once safety-relevant information is generated, sharing it while preserving legitimate confidentiality and preventing competitive harm presents distinct technical challenges. Most safety technologies also give AI labs a competitive edge, which prevents information sharing. A technical platform where these safety and security issues and solutions can be shared without leakage would be helpful. Some vulnerabilities or jailbreaking methods might be model-specific which limits the effectiveness of information sharing. A neutral clearinghouse such as AISI could facilitate anonymisation and reduce legal uncertainty for participants.

3.4 Case study: chain-of-thought sharing

Three labs (A, B, C) develop advanced reasoning models with chain-of-thought capabilities and independently identify safety concerns by displaying reasoning traces. Raw traces contain: proprietary instructions/guardrails that could be reverse-engineered; technical vulnerabilities creating jailbreaking attack surfaces; and exploitable edge cases. Labs want to share: (1) common structural standards for displaying sanitised reasoning, (2) intervention strategies for handling problematic traces, and (3) evaluation benchmarks for measuring trace quality/safety.

Chapter I prohibitions assessment. Information sharing must be assessed on whether it constitutes restriction "by object" (competitively sensitive information removing uncertainty between participants) or "by effect" (appreciable negative market impact).

Structural standards would not constitute restriction by object (no commercially sensitive strategies revealed) or by effect (standardization generally pro-competitive, encourages interoperability). *Intervention strategies* might be regarded as restriction by object (safety controls linked to commercial strategy/deployment plans) and by effect (may reduce independent decision-making on model release timing, triggering Section 2(2)(b) of Competition Act). *Evaluation benchmarks* would not qualify as restriction by object but may be restriction by effect depending on market structure.

Section 9 exemptions assessment. For exemption under Section 9, information sharing must: (1) promote technical progress/innovation, (2) allow consumers fair share of benefits, (3) not impose unnecessary restrictions, and (4) not eliminate competition.

Structural standards would likely qualify for exemption if voluntary and non-exclusive. *Intervention strategies* unlikely to qualify due to close connection to commercial strategies; labs must demonstrate information shared was reasonably necessary for pro-competitive gains. *Evaluation benchmarks* may qualify as sharing could lead to technical progress in safety measures with positive consumer pass-on.

3.5 Legal mechanisms for information sharing

Section 6 of Competition Act 1998 allows CMA to recommend to Secretary of State to establish exempt category of agreements. Where multiple similar agreements likely meet Section 9 conditions, CMA may recommend Block Exemption Order (BEO) providing ex-ante legal certainty.

The CMA's Guidance for Horizontal Agreements provides criteria: (1) Pro-competitive, (2) Necessary and proportionate, (3) Non-discriminatory (FRAND principles), (4) Aggregated or anonymised, (5) Exclude competitively sensitive information, (6) Voluntary and transparent, and (7) Third party acts as trustee.

Table 3: Comparison of legal mechanisms

Feature	**Safe Harbor** (via CMA Guidance)	**Individual Exemption**	**Block Exemption**
Legal Certainty	Medium-High: Strong comfort based on CMA enforcement intentions, but not absolute legal protection.	Low: Depends on robust self-assessment and evidence; if challenged, companies bear the burden of proof. Uncertainty is higher in novel/complex areas like AI.	High: Automatic exemption if the agreement strictly complies with all conditions. Provides highest legal certainty for compliant conduct.
Flexibility	Medium: Guidance can be principles-based but still sets defined parameters. More adaptable than a BEO.	High: Can apply to any bespoke AI safety sharing arrangement, tailored to specific needs.	Low: Agreement must conform strictly to the BEO's rules. Less adaptable to unique situations
Current Status	None for AI	Always available but high risk without specific guidance	None; requires government action

4 Institutional mechanism design

We evaluate nine policy options for incentivising information sharing, rated across feasibility, political will, and effectiveness (1-5 scale).

4.1 Economic analysis of incentives

Economic research offers quantitative insights into strategic dynamics. Gordon et al. (2003) and Gal-Or and Ghose (2004, 2005) showed that security technology investments and information sharing can function as strategic complements rather than substitutes. Strategic complementarity means that the benefit to one firm of increasing its safety investment increases when other firms increase their investments. This occurs when positive spillovers exist on either the **Demand side**—information sharing increases overall market confidence or **Fixed-cost side** —shared information reduces fixed costs (developing safety methodology, benchmarks, interpretability tools) rather than variable costs (model-specific implementation).

Gal-Or and Ghose (2005) find that benefits increase with firm size and in more competitive industries. Given AI development requires enormous capital and operates in an intensely competitive landscape, leading AI labs could derive substantial benefits from structured information sharing. In sequential, dynamic environments where companies observe others' contributions, the first firm committing to sharing moves favorably for all firms, triggering positive cascades through the ecosystem. The fixed cost channel seems more plausible for frontier AI than the demand side, as pre-training and evaluations are fixed costs, while the evidence on demand estimation remains unclear. Since security risks are not directly linked to the user, the user might not increase the overall demand much when safer products enter the market.

4.2 Why labs want (and don't want) to share

Incentives for sharing. Researchers have genuine concern for safety, motivating labs to signal commitments to attract safety-conscious talent (Odeh, 2021). Labs also cultivate regulatory goodwill through voluntary safety collaboration, moderating regulatory intervention as seen in nuclear, chemical, and aviation industries. Finally, collaboration strengthens relationships between researchers, creating infrastructure for coordinated responses during potential AI safety crises.

Disincentives for sharing. Legal ambiguity creates costly friction, as evidenced by the current practice of information sharing through lawyers. Sharing creates paper trails, exposing labs to future liability (tort claims, criminal charges). First-mover disadvantage exists where costly safety research becomes a public good benefiting competitors. Different labs have genuine disagreements over priorities, danger thresholds, and risk assessments.

4.3 Evaluation of nine mechanisms

1. Liability Shields (Feasibility (F): 1.5, Political Will (PW): 2, Effectiveness (E): 4.5): Broader shields beyond competition law protecting labs reporting to AISI from civil/criminal liability. Most effective but requires legislative amendment. Precedents exist (Public Interest Disclosure Act 1998, CISA Section 105).

2. AISI as Clearinghouse (F: 4, PW: 4, E: 4): AISI operates as a neutral third-party intermediary. AISI's legal counsel designs protocols that aggregate and anonymise information to benefit safety without enabling collusion. Requires Block Exemption Order approval but significantly reduces antitrust concerns.

3. IP Protections (F: 4.5, PW: 3, E: 1.5): Patents for safety innovations. Light-touch but limited, since patents limit the sharing of safety techniques. Royalty payments between labs would incentivise safety innovations, but seem unlikely.

4. AISI Lab Safety Scoring (F: 4, PW: 3, E: 2.5): Certification system affecting lab reputation. This may create adverse selection (labs selectively report positive information) and may damage AISI's working relationships. These create public costs for labs that don't share information and offer labs a way to get a good pubic reputation for sharing information privately.

5. Safety Taxes (F: 2, PW: 2, E: 3.5): Mandating risk mitigation when labs receive shared information creates a compliance burden, incentivising sharing to slow competitors. But this may undermine collaborative safety culture.

6. Red-Team Bounty Pool (F: Existing, PW: Existing, E: 3): AISI's existing program (£3,000-£15,000 per submission). Pricing appropriately is difficult; without complementary liability shields, labs may remain reluctant to report internally discovered vulnerabilities.

7. Public Compute (F: 5, PW: 3, E: 2): Compute access rewards. Highly feasible but limited effectiveness due to scale disparity between public and private compute of advanced GPUs and other AI chips such as TPUs.

8. Public Liability Insurance (F: 2.5, PW: 3, E: 2): Required insurance for AI harms. Insurance industry lacks expertise to price AI risk. But labs are incentivised to withhold information to maintain lower premiums.

9. Private Governance (F: 3.5, PW: 3, E: 2.5): AISI licenses private organizations as independent certifiers, creating a competitive certification market. Consumer awareness not guaranteed; potential race-to-laxest-certification.

AISI as an arms-length body. ALB status would enhance trust with labs reluctant to share with the enforcer of the upcoming UK AI bill, more credibly maintain independence, and more easily attract specialized talent. While AISI could theoretically function as a clearinghouse while remaining part of DSIT, this might face greater challenges regarding trust and perceived neutrality. Whether the status of an arm's-length body is sufficient to address Frontier Labs' concerns about sharing and exposing information with the government remains to be seen.

5 Discussion and policy implications

Our analysis reveals three critical institutional design requirements for effective AI safety information sharing frameworks. We frame these as analytical findings about necessary institutional features, though policy implications are clear.

5.1 Neutral clearinghouse institutions

Our comparative analysis demonstrates that effective safety information sharing in competitive markets benefits from neutral intermediary institutions that can aggregate, anonymise, and redistribute information while maintaining both legal compliance and technical trust. Modeled after EudraVigilance and ISAC/ISAO, which rely on trusted third parties to process sensitive information by removing commercially identifying details while preserving safety value.

In the UK, AISI possesses the key characteristics of effective clearinghouses: established technical credibility (with frontier labs) and recognized expertise, existing pre-deployment evaluation relationships, and government affiliation enabling regulatory coordination. Our analysis suggests three necessary institutional capacities must be developed to move forward: (1) **in-house legal expertise in UK competition law** to design information processing protocols maintaining compliance while maximizing safety value, (2) **collaboration with the Competition Market Authority (CMA)** to secure necessary legal frameworks, specifically block exemptions for post-R&D safety information sharing that our Chapter I analysis (Section 3.3) demonstrates currently lacks legal clarity, and (3) **technical infrastructure for sophisticated anonymisation** using privacy-preserving techniques (differential privacy, secure computation, automated sensitive content detection).

This institutional model enables graduated information flows: highly sensitive discoveries shared confidentially initially, moderately sensitive information redistributed after anonymisation, and general safety insights disseminated broadly after appropriate time delays to incentivise innovation. The clearinghouse transforms zero-sum competitive dynamics into positive-sum collaboration. It reduces transaction costs through standardized procedures, creates accountability while maintaining confidentiality, enables pattern detection across multiple reports, and provides government visibility without requiring direct regulatory intervention that might chill innovation.

5.2 Block exemption design principles

To remedy the R&D Block Exemption Order 2022 that provides insufficient coverage for post-deployment disclosures, AISI, in partnership with DSIT and the CMA, must work to establish a new block exemption. Our analysis identifies three essential design principles for the redesign:

(1) Precise scope definition—Exemptions must explicitly cover both low-CSI/high-safety information (safety evaluation plans, security protocols) and high-CSI/high-safety information (red-teaming discoveries, deployment vulnerabilities) when shared through appropriate intermediaries, based on our classification framework in Section 3.2.

(2) Fair access (FRAND) governance—Participation must remain voluntary, non-exclusive, and open to all AI model developers and providers, both frontier and emerging, as defined by the CMA. This remedies the negative effects of the "members-only" trade-association governance model that cultivates cartel creation via exclusion (Section 2.3).

(3) Transparency and adaptive governance—Block exemptions should include CMA oversight provisions, annual reporting on information sharing volumes (without disclosing confidential details), and sunset clauses ensuring exemptions remain appropriate as AI markets evolve. Five-year review periods with extension options would balance stability with adaptability.

Such block exemption would provide ex-ante legal certainty comparable to CISA's antitrust protections, encouraging deeper collaboration on risk mitigation without chilling innovation. Our economic analysis (Section 4) suggests this could catalyse positive cascades in information sharing through strategic complementarity effects.

5.3 Complementary incentive structures

Legal clarity through block exemptions is necessary but insufficient. Even with reduced antitrust risk, labs face economic barriers (first-mover disadvantage, free-riding) and psychological barriers (reputation concerns, uncertainty about reciprocity) that discourage participation. Three complementary approaches can address these remaining obstacles:

(1) Reputational mechanisms can provide low-cost signals of safety commitment. However, these must be carefully designed to avoid adverse selection. Rather than scoring labs on safety outcomes (incentivising selective disclosure of positive information only), effective systems would focus on process metrics: frequency of participation, timeliness, and comprehensiveness. This aligns incentives toward comprehensive rather than selective sharing.

(2) Resource-based incentives including compute credits, procurement preferences, or access to national datasets can help offset first-mover costs. However, effectiveness depends critically on resource value relative to private alternatives. For frontier labs with substantial private infrastructure, public compute provides limited marginal value. More promising are unique government-held resources (national datasets, procurement access) that labs cannot easily obtain through private markets. Procurement only incentivises information sharing if public procurement is a meaningful share of labs' profit, which is not the case at the moment.

(3) Experimental approaches including regulatory sandboxes and pilot programs enable testing protocols in controlled, lower-risk settings before scaling to industry-wide participation. Pilot programs with 2-3 volunteer labs sharing information on specific challenges could demonstrate viability, refine protocols, and build trust. This staged approach addresses coordination problems by enabling first movers to demonstrate value.

Effectiveness likely depends on combination rather than any single mechanism. Labs face multiple distinct barriers, and addressing only one may prove insufficient. A comprehensive approach combining legal clarity (block exemptions), institutional infrastructure (neutral clearinghouse), positive incentives (resource access), and experimental validation (pilot programs) addresses the multifaceted coordination problem.

5.4 Cross-sectoral legal architecture

Our comparative institutional analysis reveals that effective frameworks require comprehensive legal architecture addressing multiple liability categories simultaneously. The U.S. CISA model (Section 2.1) demonstrates how antitrust exemptions, liability protections, and confidentiality guarantees work synergistically. Each protection addresses a distinct barrier: antitrust exemptions address competition law uncertainty, liability protections address tort and criminal exposure, and confidentiality guarantees address reputational concerns.

The UK currently lacks an equivalent comprehensive framework. While NIS Regulations 2018 and the forthcoming Cyber Security and Resilience Bill establish mandatory reporting requirements, they do not provide the triad of legal protections necessary for voluntary peer-to-peer information sharing. Our analysis suggests that sector-agnostic legislation providing these protections for safety-critical information sharing across multiple domains (cybersecurity, AI safety, critical infrastructure) would create economies of scale in legal framework development and reduce perceptions of AI exceptionalism.

The framework design should incorporate lessons from our institutional mechanism analysis. Effective liability shields must protect good-faith disclosures from both civil claims and criminal exposure. Clear antitrust exemptions must specify covered information categories and required conditions. Confidentiality protections must guarantee that information shared through approved channels receives protection from public disclosure (FOIA exemption) and cannot be used in subsequent legal proceedings against reporting entities (similar to CISA Section 105(D)(2)).

The framework should accommodate sector-specific clearinghouses (AISI for AI, ISACs for cybersecurity, potentially future bodies for biotechnology) while maintaining consistent core legal protections across domains. This enables specialised technical expertise in each clearinghouse while providing uniform legal certainty, potentially creating network effects as experience in one sector informs best practices in others.

6 Conclusion

This paper addresses a fundamental AI governance challenge: enabling effective information sharing on safety-critical issues while preserving competitive market dynamics. Through comparative analysis of cybersecurity and pharmaceutical precedents, detailed examination of UK competition law frameworks, and systematic evaluation of institutional mechanisms, we demonstrated viable pathways for resolving this coordination failure.

Our analysis addresses three core frictions blocking AI safety information sharing. First, the lack of trusted intermediaries—resolved through our design of AISI as a neutral clearinghouse with legal expertise, regulatory partnerships, and technical anonymisation infrastructure. Second, legal uncertainty around antitrust—resolved through block exemption principles requiring precise scope definition, FRAND access, and adaptive governance. Third, insufficient participation incentives—resolved through complementary mechanisms addressing both economic barriers (first-mover disadvantage, free-riding) and psychological barriers (reputation concerns, liability fears).

Together, these institutional innovations transform zero-sum competitive dynamics into positive-sum safety collaboration. The classification framework, legal analysis, and mechanism evaluation provide reusable tools applicable beyond the UK context. By demonstrating how thoughtful institutional design can enable safety coordination without facilitating anticompetitive behaviour, this work charts a path for AI governance that balances innovation, competition, and public safety.

References

[1] Ball, D. W. (2025). A Framework for the Private Governance of Frontier Artificial Intelligence. arXiv preprint arXiv:2504.11501.

[2] Competition and Markets Authority (2024). AI Foundation Models: Initial Report. April 2024.

[3] European Commission (2023). Guidelines on Horizontal Cooperation Agreements. Official Journal of the European Union.

[4] European Medicines Agency (n.d.). About Us. Available at: https://www.ema.europa.eu/en/about-us (accessed October 2025).

[5] European Medicines Agency (n.d.). Pharmacovigilance Overview. Available at: https://www.ema.europa.eu/en/human-regulatory-overview/pharmacovigilance-overview (accessed October 2025).

[6] Executive Order 13691 (2015). Promoting Private Sector Cybersecurity Information Sharing. Federal Register, 80 FR 9349. Available at: https://obamawhitehouse.archives.gov/the-press-office/2015/02/13/executive-order-promoting-private-sector-cybersecurity-information-shari

[7] Faure, M. (2014). Economic Analysis of Tort Law. In Law and Economics, pp. 165-202. Routledge.

[8] Gal-Or, E., & Ghose, A. (2004). The Economic Incentives for Sharing Security Information. Information Systems Research, 16(2), 186-208.

[9] Gal-Or, E., & Ghose, A. (2005). The Economic Consequences of Sharing Security Information. Economics of Information Security, 12, 95-105.

[10] Gordon, L. A., Loeb, M. P., & Lucyshyn, W. (2003). Sharing Information on Computer Systems Security: An Economic Analysis. Journal of Accounting and Public Policy, 22(6), 461-485.

[11] Network Advertising Initiative (2000). About the NAI. Available at: https://thenai.org/about-the-nai-2/ (accessed October 2025).

[12] Odeh, M. (2021). Employee Retention Through Authentic Organizational Commitment. Journal of Organizational Psychology, 21(4), 89-103.

[13] Piscitello, D. (2010). Conficker Summary and Review. ICANN. Available at: https://www.icann.org/en/system/files/files/conficker-summary-review-07may10-en.pdf

[14] Presidential Decision Directive 63 (1998). Critical Infrastructure Protection. The White House. Available at: https://irp.fas.org/offdocs/pdd/pdd-63.htm

[15] UK Competition Act 1998 (Research and Development Agreements Block Exemption) Order 2022. SI 2022/456.

[16] UK Competition Act 1998 (1998). Section 36: Financial Penalties. Available at: https://www.legislation.gov.uk/ukpga/1998/41/section/36 (accessed October 2025).

[17] UK Network and Information Systems Regulations 2018. SI 2018/506.

[18] U.S. Cybersecurity Information Sharing Act of 2015. Pub. L. No. 114-113, 129 Stat. 2242.

[19] U.S. Department of Justice & Federal Trade Commission (2014). Antitrust Policy Statement on Sharing of Cybersecurity Information.

A Detailed analysis of institutional mechanisms

A.1 Liability shields

Feasibility (1.5/5): Requires lobbying for amendment to UK legislation. Block exemption creates liability exemptions for competition law but not for other liabilities (torts, criminal charges). Precedents: Public Interest Disclosure Act 1998 (whistleblower protection); CISA Section 105(c)/(d) requires government keep shared information secret; German Lieferkettensorgfaltspflichtgesetz where due diligence prevents automatic liability.

Political Will (2/5): Some political appetite exists. While introducing new legislation faces frictions, it is in UK's interests to remain relevant in global AI development. Having foremost safety research institution with collaborative relationships with leading labs is key advantage. Liability protections deepen relationships; AISI's research talent helping technically helps UK get closer to US (desired geopolitical goal).

Effectiveness (4.5/5): Directly addresses common barrier. If AISI remains subject to oversight by government regulator, labs may fear sensitive information could be disclosed or misused. Liability shield avoids this problem. AISI can work with lab to help resolve issues. Second-order effect: cultivates trust, incentivising further sharing.

Concerns: Raises questions of fairness for potential AI damage victims who cannot pursue legal remedies. Potential moral hazard where waiving legal consequences reduces incentives for proactive safety investment. However, we envision this refers to proactive sharing during internal development/testing (pre-deployment), so unlikely to be real external harms.

A.2 AISI as information clearinghouse

Feasibility (4/5): Key challenge is first needing Block Exemption Order approved by CMA and Secretary of State. While we anticipate broad support, this represents significant administrative effort. Requires hiring specialized competition law counsel and developing robust information filtering/anonymisation systems.

Political Will (4/5): Strong appetite anticipated. Gains all political advantages of Liability Exemptions with minimal downside. More appealing if AISI not playing direct role in assisting frontier lab technical research, making labs less dependent on foreign countries.

Effectiveness (4/5): Main advantage is government-attached neutral intermediary significantly reduces antitrust concerns. AISI's legal team creates standardized procedures, removing commercially sensitive details while preserving safety value. Companies want to share, but antitrust consequences are severe. Labs could rely on AISI's well-designed process rather than navigating complex competition law individually.

Professional independence maintained: AISI's legal counsel advises solely AISI on clearinghouse operations while participating labs retain own independent counsel. This ensures compliance while reducing legal uncertainty without creating attorney-client relationships compromising clearinghouse neutrality.

A.3 Economic analysis expanded

Strategic complementarity means benefit to one firm of increasing sharing/investment increases when other firms increase theirs. However, this only occurs when positive spillovers exist on demand side (information sharing increases overall market confidence, expanding demand for all firms) or fixed-cost side (shared information reduces fixed costs like developing safety research methodology, benchmarks, interpretability/alignment tools rather than variable costs like model-specific implementation, per-query safety techniques).

Without spillover effects, free-riding dominates. With complementarity, increase in one firm's sharing induces others to increase theirs, creating virtuous cycle. Particularly relevant to AI: Gal-Or and Ghose (2005) find benefits increase with firm size and in more competitive industries. Given AI development requires enormous capital and operates in intensely competitive landscape, leading labs could derive substantial benefits from structured sharing.

While models examine single-shot simultaneous decisions, in sequential, dynamic environments like AI development where companies observe others' contributions, conclusion is more, not less, sharing. First firm committing to sharing moves favorably for all firms, convincing entrant to follow suit, triggering positive cascades. This improves opportunities for tacit collusion as increased sharing and technology investment lead to less aggressive price competition, benefiting all participants with strictly higher profits under sequential than simultaneous dynamics.

A.4 Qualitative incentives and disincentives

Why labs want to share: Researchers have genuine safety concern. This motivates labs (commercial entities) to signal safety commitments to attract safety-conscious AI talent. Odeh (2021) found authenticity of commitments critical in retaining motivated staff. Labs voluntarily signal safety care through collaboration to cultivate regulatory goodwill, discouraging tighter regulation. Precedents in nuclear, chemical, aviation demonstrate how voluntary safety collaboration moderates regulatory intervention. However, this requires industry-wide collaboration—classic coordination problem where expected policy gains alone are insufficient incentives beyond countervailing concerns. Labs want collaborative environment inspiring costly safety research; if shared as public good, others free ride. Finally, collaboration strengthens pre-existing relationships, setting up formal networks providing vital infrastructure for coordinated responses during potential AI safety crises.

Why labs don't want to share: Legal ambiguity around information sharing between labs. Many safety-relevant information kinds have less legal certainty regarding anti-competition laws. Source at frontier lab confirmed current sharing is done via lawyers. In many cases, information like internal red teaming results not even useful to share (not universal). Cost/usefulness asymmetry; costly and inconvenient process introducing collaboration frictions.

Sharing information creates paper trail exposing labs to future liability. While we explored legal carve-outs under competition law, these don't shield from other liability forms (tort claims, criminal charges). Legal liability often hinges on what company knew; disclosure makes establishing knowledge easier. Keeping information internal reduces legal action likelihood since potential claimants may lack evidence to justify cases.

More inherent reason: first mover disadvantage where safety information, particularly techniques, are costly to develop yet become public good when shared. Competitors access without granting sharer explicit reciprocal benefit. We get unsustainable dynamics when not positive demand spillovers (consumer demand doesn't meaningfully increase with safety) or if safety sharing doesn't reduce fixed costs (or reduces only variable costs).

Different researchers and labs have genuine disagreements over priorities, danger thresholds, risk assessments. Differing standards over what constitutes meaningful safety concerns or acceptable risk means one party may think sharing particular information unnecessary. Finally, certain information kinds have minimal sharing incentive (specific model vulnerability exploitable). Addressing requires restricting model access and pausing development, both costly.

Table 4: Cybersecurity Information Sharing Act (CISA) provisions

Provision	Allows/Requires	Legal Protections	Incentives
104(C)	Share cyber threat indicators with each other/government	Permitted for cybersecurity purposes	Voluntary collaboration
104(D)(1)	Security controls protecting against unauthorized access	Appropriate controls implemented	Shared info remains secure
104(D)(2)	Scrubbing personal data before sharing	Info anonymised for privacy	Balances transparency with privacy
104(E)	Sharing threat/defense info not anti-competitive	Antitrust exemption	Removes fear of legal action
105(D)(2)	Government-shared info protected as proprietary	FOIA exemption	Enables trust in government collaboration
106(B)	Entities sharing under 104(C) protected from antitrust liability	Liability protections	Reduces legal risk

B CISA provisions

C Additional case study details

C.1 Pharmaceuticals extended

Pharmaceutical sector benefits from industry-led trade associations (IFPMA, PhRMA) collaborating on global standards and best practices. Broader knowledge-sharing occurs at Drug Safety Symposium, CDISC Interchange, Reuters Events Pharma discussing pharmacovigilance methods, data standards, safety governance.

Legal safe harbors exist: in US, companies reporting problems through official channels when completing acquisitions are generally protected from some litigation consequences. As of 2018, EudraVigilance allows companies to see industry-wide safety trends without exposing which competitor contributed which report, preserving confidentiality and reducing duplicative reporting.

Pharma companies want to avoid negative publicity, so key challenge is ensuring sharing safety data doesn't become competitive disadvantage. Sector addresses this by aggregating and anonymising data. Adverse reaction reports (ICSRs) created by EMA don't name companies publicly; they feed into broader safety signals. When serious risk identified, all similar drug manufacturers alerted simultaneously. No single company unfairly singled out at early stage—focus is class-wide safety.

Industry engaged in pre-competitive collaborations. Firms pooled data on drug toxicology and early clinical trials to improve safety assessments for all, under consortia agreements protecting proprietary details. Collaborations with neutral coordinator (public-private partnership, professional association) help each company learn from others' failures without fearing immediate commercial fallout.

C.2 Digital advertising extended

Frontier Model Forum (FMF), founded by OpenAI, Anthropic, Google DeepMind, Microsoft, later joined by Amazon and Meta, has emerged as most visible framework for coordinating AI safety information sharing (March 2025). FMF positions itself as facilitator of cross-firm communication on safety risks, organizing efforts around: "vulnerabilities, weaknesses, and exploitable flaws;" "threats;" "capabilities of concern." However, FMF currently limited to members, leaving smaller/independent AI companies excluded.

FMF rise reflects common trend where trade associations lead early coordination before regulation catches up. Useful parallel: NAI helped set standards in online data collection later codified into law. While NAI not ultimately successful delivering strong, lasting privacy protections (compared to GDPR), it remains valuable case study in early self-regulation. Illustrates how industry facing political pressure and low public trust can coordinate to delay or shape upcoming regulation.

As voluntary initiative, NAI avoided antitrust pitfalls by limiting collaboration to privacy/consumer protection, not pricing or market division. FTC endorsement lent legitimacy, helped delay formal regulation. While no explicit legal immunity existed, transparency and public standards focus kept initiative within legal bounds. By agreeing baseline privacy rules, firms aimed preventing race to bottom for aggressive data practices provoking sector-wide backlash.