The problem

Frontier AI systems can pose serious risks if misused. UK users can access them directly via overseas APIs (e.g. ChatGPT, DeepSeek) without meeting any UK safety standard. Companies may choose to limit access voluntarily (e.g. OpenAI's Sora rollout), but no UK law currently blocks a high-risk AI unless embedded in a regulated service or a domain-specific regime like medical devices.

The core gap is **interpretability**: the ability to understand how an AI makes decisions. Without it, dangerous capabilities or deception may go undetected. This is crucial for systems with **agency** (autonomous task initiation), **recursion** (self-modification or spawning agents), long-horizon planning, or power-seeking tendencies — all tied to catastrophic risk and absent from current UK/EU law.

The goal

The UK hosts leading AI institutions (DeepMind, Anthropic, Oxford, Cambridge, Imperial), AISI, and is culturally close to US firms. If the US slows AI via domestic rules, the UK could become a relocation target — making our standards globally influential. UK GDP (~$3.3T) plus the EU market matches the US; tying access to interpretability benchmarks would create strong compliance incentives.
This project would design a **legal and technical "gate"** so no AI above a high-risk threshold could be offered to UK users without passing an **Interpretability Readiness Level (IRL) certificate**. Thresholds would align with the EU's "systemic-risk GPAI" definition but also require evidence for systems showing agency, recursion, delegation, or cross-task goal retention.
The gate would be created via a short enabling Act of Parliament, with detail in a **Statutory Instrument** (SI).

IRL certificate elements

- **Risk-relevant coverage**: causal control over high-risk components, including agency/recursion subsystems.
- **Deception checks**: oversight transformations, sandbag detection, irrelevant-task canaries.
- **Robustness**: behaviour under distribution shifts.
- **Anti-gaming**: withheld suite rotation, multi-party red-teaming, perturbations.
- **Access for testing**: read-only secure-enclave examination of parameters.

If evidence were incomplete, unreliable, or gameable, certification would be refused; providers could only serve the UK by downgrading below scope or geofencing.

Enforcement

- **Service-restriction orders** for intermediaries (app stores, API marketplaces, payments, search indexing).
- **Access-restriction orders** for ISPs as a last resort.

EU alignment
Designed for mutual recognition with systemic-risk GPAI rules, with a crosswalk annex; agency/recursion addressed as extra interpretability requirements, not a separate category.

Framing to avoid perception of “more UK internet restrictions”
Tightly scoped to the highest-risk systems; no powers over lawful content or ordinary services; confidential audits by accredited third parties; statutory three-year review.