Unlocking Capability Guarantees through AI Procurement

Afek Shamir
Independent Researcher
London, United Kingdom
afek.shamir@gmail.com

Abstract

As middle powers integrate frontier AI models from a concentrated number of private developers, they risk strategic dependencies that may undermine sovereignty, security, and continuity of critical services. This paper proposes adapting a well-established defence procurement tool – capability guarantees, defined as contractual rights ensuring adequate operational sovereignty and risk mitigation – to the procurement of frontier AI. By acting as coordinated consortia, these AI middle powers – states with significant economic, diplomatic, security, or regulatory advantages but high dependency on foreign AI providers – could leverage public procurement to enhance de facto international standards for AI safety and security. The approach can complement hard regulation by creating market incentives for safer AI deployments.

1 Introduction

The rapid proliferation of powerful general-purpose AI models has concentrated capability and operational control in a handful of private developers. Unlike traditional software, the opaque nature of frontier AI models and the speed with which their capabilities are progressing makes it challenging to verify that AI models' safety, security, and alignment with human preferences is guaranteed post-deployment.[1] This combination – concentration in foreign private hands, opacity that prevents independent verification, and rapid capability progression – can create acute strategic dependencies for states. In particular, states which increasingly rely on AI systems for critical applications (bar some that host leading AI Safety and Security Institutes) cannot independently assess their safety and security, cannot guarantee continued access to frontier capabilities, and cannot reliably predict how capabilities will evolve and what risks they should assess for and mitigate. Policymakers in AI middle powers (e.g. UK, France, Germany, Canada, Japan) therefore face a strategic choice: accept dependence on foreign providers to reap benefits from AI adoption, or insist on forms of AI adoption that may restrict utility and economic growth.

This paper proposes a different option: using public procurement as a strategic governance tool to secure capability guarantees while maintaining access to frontier AI capabilities. The argument proceeds in four steps. First, we establish that the public procurement process – especially when coordinated across aligned states – can serve as a forum to negotiate contractual commitments for verifiable operational guarantees that protect national interests. Second, we draw on the precedent of 'capability guarantees' in defence procurement, particularly the F-35 Joint Strike Fighter program, to demonstrate how purchasing states have historically managed strategic dependencies on foreign-supplied critical technologies. We define capability guarantees as contractual terms – distinct from technical specifications – designed to manage strategic dependencies and ensure a purchasing state retains a sufficient degree of operational sovereignty and capacity to verify safe and secure integration.

[1] For data on the speed of change in model capabilities, see [11]. For difficulty to guarantee model safety, security, and alignment, see [9].

Third, we adapt these principles to AI procurement, identifying specific security vulnerabilities that middle powers face and proposing corresponding contractual mechanisms. Finally, we explore the feasibility of a middle power buyers' consortium, analysing both the economic leverage such coordination would provide and the institutional challenges it must overcome.

2 Outlining the Precedent: Capability Guarantees in Defence Procurement

Regulation and public procurement are complementary but distinct governance levers. Regulation (for example the EU AI Act) establishes a set of rights, obligations, and prohibitions that apply across the market. Procurement, by contrast, is a targeted, transaction-level instrument that can create incentives for raising governance standards through contractual terms, purchasing specifications, and operational requirements. Practically, regulation shapes what vendors must or must not do to operate in a market, while public procurement shapes what vendors choose to offer to win public contracts.

In defence contexts, procurement-based governance becomes important when critical capabilities are acquired from foreign suppliers. Unlike civilian technologies, defence systems often involve classified information and can be sufficiently critical for service interruptions to materially threaten national security.[2] In such cases, standard commercial procurement may be insufficient. States often require capability guarantees: contractual covenants bundled into procurement agreements that go beyond performance specifications and confer the buyer explicit, enforceable rights that preserve operational sovereignty over a purchased capability.

The F-35 Joint Strike Fighter program serves as a particularly fitting case study to the case of frontier AI, because it involves the procurement of highly complex, foreign-supplied (for those outside the U.S.), albeit not general-purpose, technology essential for national security.[3] The program's procurement agreements were structured to provide partners – which included the UK, Italy, Netherlands, Australia, Canada, Norway, Denmark, and others – with varying degrees of operational sovereignty by setting contractual terms with the U.S. government and the relevant private provider, Lockheed Martin.[4] Specifically, Tier 1 partners, such as the UK, secured rights beyond the purchase of an aircraft.[5] These capability guarantees included access to technical data and source code, capacity to independently test, maintain, and configure the fleet, and the establishment of some domestic maintenance, repair, and overhaul facilities in the UK. Given the UK's stated goal to achieve "freedom of action",[6] securing its privileged partner status was a pathway to acquire its F-35 fleet without overreliance on the original equipment manufacturer (Lockheed Martin) or the U.S. government.[7]

Despite being designated a Tier 1 partner, a recent UK National Audit Office (NAO) report described that "freedom of action" was not fully achieved. Some cited reasons include significant delays in integrating sovereign weapons systems (the Spear 3 missile) and an overreliance on providers to deliver spare parts at the required rate.[8] Nevertheless, the core principle remained: the procurement of a critical capability can be bundled with the guarantees necessary to ensure its safe and secure operation by the country deploying the capability. An effective use of capability guarantees in frontier AI procurement – involving continuous, active management of sovereign goals – can complement regulatory regimes by creating additional commercial value for compliance.[9]

3 Applying the Precedent to AI Procurement

The rationale of capability guarantees in defence can be transferable to the procurement of frontier AI. When a government procures an AI model for critical applications – as seen in the UK's recent agreement with major AI labs like Anthropic or the US Department of Defence's investments in

[2] [16], [4], and [21]
[3] [2]
[4] [13]
[5] [7]
[6] Defined in [17] as the ability to use the capability at a time and place of its choosing.
[7] [17]
[8] [17]
[9] [24]

Anthropic, Google, OpenAI and xAI[10] – it is acquiring infrastructure that holds tremendous risks which must be routinely identified and mitigated. These risks stem from AI systems' characteristics. AI's opacity makes verification challenging, its rapid capability progression means it is challenging to transparently evaluate and test their behaviour, and its general-purpose nature means a single system can be repurposed for tasks beyond its initial design. Therefore, without robust capability guarantees, middle powers could expose themselves to security vulnerabilities, including but not limited to:

- **Strategic deception and misalignment:** A frontier model could be deceptively aligned by design, or exhibit "scheming."[11] That is, an integrated AI model might appear to function as intended during evaluations only to pursue misaligned goals or secret loyalties once integrated into sensitive systems. This could happen, for example, through malicious data poisoning.[12] For a state, this is a security risk that could manifest in adverse outcomes (e.g. an intelligence analysis subtly manipulating methodologies or summaries to favour a particular geopolitical outcome).
- **Loss of data sovereignty:** Procured models, particularly those accessed via APIs, could become conduits for sensitive state data.[13] Without guarantees on data handling and processing, a middle power might have insufficient assurance that classified intelligence or critical information is not being logged, stored, or analysed by the foreign provider for purposes beyond the contractual scope.
- **Discontinuation of service:** Arguably, the most acute economic and security risk is the "kill switch." In a geopolitical crisis, the host government of the AI developer could compel it to restrict or terminate access for a middle power. If a nation's emergency response or defence readiness relies on that service, such a denial would be a critical operational issue.

4 Creating a Buyers' Consortium for Frontier AI

While a single middle power may lack the market leverage to impose significant demands on large AI developers, a coordinated approach among states like the UK, Canada, Japan, and large EU states would likely be sufficient to set a baseline minimum contractual agreement regime.[14] While precise figures for government AI procurement remain obscured due to the novelty of this enterprise and its associated confidentiality, approximate estimates suggest that these states' combined procurement power would likely be a significant market force.[15] Such a consortium could reasonably demand some capability guarantees as standard contractual covenants in public sector contracts purchasing frontier AI capabilities. Reflecting the guarantees sought in the defence precedent, these may include (but not limited to):

- **Mandated access for independent testing:** Procurement contracts could stipulate rights for a trusted national entity (e.g., a state-backed AI Safety or Security Institute) to conduct rigorous evaluations of the model's reliability and governability.[16] This may include access not just to the model's outputs, but to the necessary technical documentation, evaluation tools, and compute resources to specifically test for behaviours like oversight subversion and sandbagging.[17]

[10] [10] and [6]

[11] See [15], [1], and [20]

[12] [22]

[13] [18]

[14] [5]

[15] Rationale: If we use some conservative assumptions, we can attempt to estimate the collective procurement power of AI middle powers. In 2025, the UK's procurement of AI models is approximately £3.3 billion as per [25] UK Public Sector AI Procurement Tracker. If we assume similar per-capita spending adjusted for GDP for major EU economies (France, Germany, Italy, Spain, Netherlands) and other middle powers (Canada and Japan), we can estimate these figures to be €8-12 billion and $3-5 billion, respectively. Together, the final figure is roughly €15-20 billion in 2025. Applying the market growth rate of 36.89% annually as per [23] estimate, this current figure means these states could collectively represent €72-96 billion in annual AI procurement by 2030. Given AI is an exponentially improving technology – and prone to winner-takes-most dynamics – these figures could be far greater and are thus conservative estimates.

[16] [19]

[17] [19]

- **Transparent submission of 'safety cases':** Procurement processes could demand that the control measures themselves are transparent and can be independently audited through the submission of a safety case: that is, assessable arguments for the safety of a system in a given context.[18] For instance, an AI developer could go beyond solely conducting a monitoring process to detect unintended behaviour, but go further to provide auditable evidence of the effectiveness of risk assessments and mitigations, potentially also allowing the buyer to red-team the control itself.[19]
- **Tiered operational sovereignty:** The demand for model weights is a significant barrier for developers. A more feasible approach is to contract for tiered levels of operational control based on the criticality of the application, mirroring the 'Tier 1' partner status in the F-35 program. For low-stakes tasks, standard API access may suffice. For use in critical infrastructure or defence, the contract could mandate more robust guarantees, such as on-premise deployment. Across the board, assurances against arbitrary or sudden termination of service might be requested to avoid significant economic and security risks.

5 Conclusion and Areas for Future Research

In sum, this paper argues that middle powers can mitigate some security risks of AI dependency by transforming public procurement from a buyer-seller transaction into a strategic tool for governance. By adapting the precedent of capability guarantees from defence and acting as a unified buyers' consortium, middle powers may be able to demand auditable assurance of a model's reliability and governability, thereby safeguarding sovereign interests and security.

There is ample space for future research in this direction. For one, this procurement-led governance model is not without challenges: the market power of frontier AI labs is immense, and their primary government partners remain the U.S. and Chinese governments. Second, the approach of setting standards via procurement must complement, not replace, hard regulation. Procurement can act as a tool for norm diffusion that can initiate a ripple effect, incentivising vendors to elevate standards for the AI ecosystem as a whole – but governments have thus far not been the primary customer of frontier AI.[20] Thirdly, the deep military and intelligence integration of many middle powers with the U.S. (e.g., through NATO or the Five Eyes) may practically limit their ability to impose divergent standards on systems intended for joint operational use. Future work should focus on developing standardised contractual language for these guarantees and designing the institutional frameworks for shared AI evaluation and auditing.

References

[1] Apollo Research. 2025. "More Capable Models Are Better at In-Context Scheming." Apollo Research. June 19, 2025.

[2] BAE systems. 2025. "What Is the Joint Strike Fighter (JSF) Program?" BAE Systems | United States. 2025.

[3] Balesni, Mikita, et al. 2024. "Towards Evaluations-Based Safety Cases for AI Scheming." ArXiv.org. 2024.

[4] Barrie, Douglas. 2025. "Europe's Air of Dependence." IISS. 2025.

[5] Bradford, Anu. 2019. *The Brussels Effect: How the European Union Rules the World*. Oxford University Press.

[6] CDAO. 2025. "CDAO Announces Partnerships with Frontier AI Companies to Address National Security Mission Areas." Chief Digital and Artificial Intelligence Office. July 14, 2025.

[7] Chapman, Bert. 2018. *Global Defense Procurement and the F-35 Joint Strike Fighter*. Palgrave Macmillan US.

[18] [12]
[19] [14] and [3]
[20] [8]

[8] Dor, Lavi M. Ben, and Cary Coglianese. 2021. "Procurement as AI Governance." IEEE Transactions on Technology and Society 2 (4): 192–99.

[9] DSIT. 2025a. "International AI Safety Report 2025." GOV.UK. January 29, 2025.

[10] DSIT. 2025b. "Memorandum of Understanding between UK and Anthropic on AI Opportunities." GOV.UK. February 14, 2025.

[11] Epoch AI. 2025. "Machine Learning Trends." Epoch AI. April 11, 2025.

[12] Hilton, Benjamin, et al. 2025. "Safety Cases: A Scalable Approach to Frontier AI Safety." ArXiv.org. 2025.

[13] House of Commons Defence Committee. 2017. "HC 326 Unclear for Take-Off? F-35 Procurement Second Report of Session 2017-19."

[14] Irving, Geoffrey. 2024. "Safety Cases at AISI." AI Security Institute. 2024.

[15] Meinke, Alexander, et al. 2024. "Frontier Models Are Capable of In-Context Scheming." ArXiv.org. 2024.

[16] Mejino-López, Juan, and Guntram B. Wolff. 2025. "Europe's Dependence on US Foreign Military Sales and What to Do about It." Bruegel. June 20, 2025.

[17] NAO. 2025. "The UK's F-35 Capability" National Audit Office (NAO). July 10, 2025.

[18] NextLabs. 2025. "Securing Sensitive Data in AI Models: A Critical Priority in Large Training Data Sets - NextLabs." August 20, 2025.

[19] Pistillo, Matteo, and Charlotte Stix. 2025. "Assurance of Frontier AI Built for National Security." ArXiv.org. 2025.

[20] Schoen, Bronson, et al. 2025. "Stress Testing Deliberative Alignment for Anti-Scheming Training." ArXiv.org. 2025.

[21] Soler, Paula. 2025. "Defence Industry Warns EU to 'Urgently' Curb Dependence on Rare Earths." Euronews. October 14, 2025.

[22] Souly, Alexandra, et al. 2025. "Poisoning Attacks on LLMs Require a Near-Constant Number of Poison Samples." ArXiv.org. 2025.

[23] Statista. 2025. "Artificial Intelligence - Global | Statista Market Forecast." Statista. March 2025.

[24] Tomei, Philip Moreira, Rupal Jain, and Matija Franklin. 2025. "AI Governance through Markets." ArXiv.org. 2025.

[25] Tussell. 2025. "UK Public Sector AI Procurement Tracker." Tussell.com. 2025.