Calibrating Oversight for Agentic Frontier Models

ACM Europe Technology Policy Committee (ETPC)

April 27, 2026

Dr. Patricia Alves, Dr. Michael Beaudouin – Lafon, Gaston Besanson, Dr. Luigi Di Biasi, Dr. Chris Hankin, Francisco Medeiros, Dr. Eirini Ntoutsi, Tom Romanoff, Alejandro Saucedo, Gerhard Schimpf, Serafeim Triantafyllou

Subject: Sustaining Technical Integrity in the Age of Agentic AI

1. Executive Summary and Methodological Note

The ACM Europe Technology Policy Committee (ETPC) submits this analysis to assess how the EU Digital Omnibus package put forward by the European Commission (COM(2025)0837), including the AI-specific proposal amending the AI Act (COM(2025)0836), could respond to the rapidly improving cyber-relevant capabilities of advanced agentic AI systems. Recent disclosures regarding **Claude Mythos Preview**[1], supplemented by public evaluations from the **UK AI Security Institute (AISI)**[2], indicate an advancement in cyber-relevant performance of AI systems. These disclosures support concerns that current oversight approaches may become misaligned with the speed and form of capability change.

While we support the European Commission's intent to simplify the regulatory burden, we warn that **the sequencing of implementation must not create gaps in technical oversight**. This document argues that simplification must be paired with robust technical oversight for high-capability systems[3].

This analysis relies on vendor-reported benchmarks and independent evaluations indicating improved proficiency in multi-step cyber-attack simulations. Accordingly, the policy recommendations contained herein are framed as precautionary and capability-sensitive, addressing the potential for systemic risk identified in controlled evaluations rather than responding to documented large-scale misuse.

2. Observed Developments: Agentic Capabilities and Cyber Relevance

Technical analysis of some frontier systems[4] suggests a transition toward models capable of autonomous problem-solving in complex software environments.

- **Expert-Level Performance in Technical Tasks:** Anthropic reported a 93.9% score

[1] Anthropic. 2026. *System Card: Claude Mythos Preview*. Anthropic. Retrieved April 27, 2026 from https://www-cdn.anthropic.com/08ab9158070959f88f296514c21b7facce6f52bc.pdf

[2] AI Security Institute (AISI). 2026. Our evaluation of Claude Mythos Preview's cyber capabilities. Retrieved April 27, 2026 from https://www.aisi.gov.uk/blog/our-evaluation-of-claude-mythos-previews-cyber-capabilities

[3] This refers to General-Purpose AI (GPAI) with systemic risk (Article 51 EU AI Act)

[4] We distinguish between the core technical artifact (Model) and the operational environment (System) where it is integrated with tools, agentic loops, and human scaffolding.

on SWE-bench[5], together with public AISI findings of continued improvement in capture-the-flag challenges and significant improvement in multi-step cyber-attack simulations, suggesting that frontier models are approaching expert-level performance in selected software engineering and cyber-relevant tasks. In particular, AISI reported first-time full completion of its 32-step 'The Last Ones' enterprise-network attack range on 3 of 10 attempts, indicating that frontier systems are beginning to sustain autonomous performance across extended offensive workflows.

- **Autonomous Vulnerability Discovery:** Anthropic reported the discovery of long-standing security flaws, such as the 27-year-old integer overflow in OpenBSD[6]. This indicates that agentic systems can navigate low-level codebases with minimal human scaffolding. Anthropic further reported that the OpenBSD finding was obtained at a low per-run cost and within a relatively modest overall campaign budget[7], reinforcing concerns that AI systems may compress not only discovery timelines but also the economic barriers to advanced vulnerability research.
- **Historical Analogue:** This shift echoes the earlier ambition of the DARPA Cyber Grand Challenge to automate vulnerability discovery and evaluation at machine speed, but now arises in the context of a more general-purpose frontier model, potentially broadening both defensive utility and the pathways to misuse. We note that the community also acknowledges that many of these vulnerabilities can already be identified by existing models such as GPT-4.5, Opus, and Sonnet. Although Anthropic's report may involve marketing efforts, the fact that some of these exploits are available makes it clear that the industry is experiencing a seismic shift in cybersecurity due to AI capabilities.
- **Compressed Exploitation Windows:** The realisation of these emerging cybersecurity gaps suggests that the discovery-to-exploitation window may shorten sharply as offensive discovery becomes increasingly automated, potentially challenging traditional patch-management cycles.

3. Governance Implications: Implementation Dependencies and Oversight Gaps

The successful implementation of the **EU AI Act** depends on its seamless integration with the broader Union cybersecurity framework, including the **NIS2 Directive** and the **Cyber Resilience Act (CRA)**. While the **EU Digital Omnibus** package aims to facilitate this integration through simplification, the ETPC identifies several critical implementation dependencies:

- **Risk of Oversight Lag:** Because implementation of certain high-risk AI obligations, in practice, depends on harmonized standards and related guidance, the current sequencing risks creating delayed or uneven oversight for frontier systems. This is particularly concerning during the current period of rapid capability acceleration.
- **Insufficiency of Static Compliance:** The agentic nature of recent models demonstrates the limits of static, documentation-first compliance when not complemented by continuous technical assurance. Relying on periodic self-

[5] Anthropic. 2026. *System Card: Claude Mythos Preview*. Anthropic. Retrieved April 27, 2026 from https://www-cdn.anthropic.com/08ab9158070959f88f296514c21b7facce6f52bc.pdf

[6] Anthropic. 2026. Project Glasswing: Securing critical software for the AI era. Retrieved April 27, 2026 from https://www.anthropic.com/glasswing

[7] Anthropic reported a per-run cost below $50 and a total campaign cost of approximately $20,000 for the OpenBSD research campaign. See: Anthropic, "Project Glasswing: Securing critical software for the AI era," April 2026.

assessment may fail to capture the dynamic risks associated with models capable of autonomous goal-seeking.

- **Proportionality, Privacy, and Trade Secrets:** Provisions protecting trade secrets and personal privacy must be balanced against the need to protect public safety. A proportionate oversight regime, incorporating specific safeguards against regulatory overreach, should ensure that competent authorities access only the safety-critical telemetry and technical documentation necessary to verify claims. This process must adhere to strict data minimization, confidentiality, and purpose limitation to prevent the compromise of proprietary IP or user privacy.

4. Policy Recommendations

To ensure a resilient implementation of the AI Act, this draft recommends exploring the following measures:

A. Risk-Triggered Continuous Technical Assurance

The ETPC proposes exploring a Risk-Triggered Continuous Technical Assurance architecture, defined here as a technical framework for automated, threshold-based reporting. This framework is a step closer to a system grounded in harmonized technical standards, such as those developed by CEN/CENELEC in support of the AI Act, in which providers of designated frontier systems report narrowly scoped telemetry to the EU AI Office.

- **Definition of "Risk":** In this context, "risk" refers to Technical Capability Risk: specifically, when a model's performance on validated benchmarks exceeds safety and security thresholds, or when the model environment is vulnerable to exfiltration, weight theft, or malicious fine-tuning. This distinguishes the trigger from the AI Act's application-based risk or traditional cybersecurity exposure.
- **Adaptive Benchmarking and Proportionality:** The architecture is capability-triggered: it is activated only when specific technical thresholds are crossed. To remain relevant as model capabilities evolve, these thresholds must be governed by adaptive benchmarking overseen by independent bodies (e.g., the EU AI Office or designated scientific institutes). This ensures the framework can respond to emerging risks such as orchestrated multi-agent attacks or advanced autonomous reasoning.
- **Scope and Supply Chain Integrity:** Telemetry should focus on safety-critical signals (ensuring the system operates as intended) and security-critical indicators (detecting adversarial exploitation, autonomous offensive workflows, or supply chain risks such as hardware provenance and SBOM anomalies). This reporting must prioritize data minimization, focusing on behavioral signals rather than security-sensitive technical assets such as model weights, which constitute core intellectual property and pose an exfiltration risk.
- **Security of the Reporting Layer:** Any such reporting architecture would require secure technical safeguards, including cryptographic attestation or comparable trusted-execution mechanisms, to protect sensitive capability information and reduce the risk of breaches in the oversight channel itself.

B. Secure Access to Supervisory Artifacts

To address the challenges posed by the opaque nature of agentic systems, this analysis recommends that oversight for designated frontier systems include secure access to logs,

evaluation artifacts, safety telemetry, and technical documentation pertaining to safety-critical capability thresholds.

- **Legal Alignment**: This requirement builds on the Technical Documentation obligations set out in Annex IV of the AI Act, ensuring that regulators have the necessary evidence to conduct meaningful technical audits.
- **IP Protection**: Access should be managed through secure, restricted protocols to ensure that supervisory transparency does not compromise the provider's core intellectual property or trade secrets.
- **Auditability**: By focusing on “capability thresholds”, the oversight remains proportionate; regulators only "deep-dive" into the specific technical artifacts that explain how a model’s offensive capabilities are being mitigated.

C. Strategic "Defensive Uplift"

The AI Act’s sandbox framework should be used more actively to support defensive security applications in Europe. In particular, Article 57 sandboxes and related testing arrangements could help critical-infrastructure operators, public-interest security teams, and trusted providers evaluate frontier systems for assisted vulnerability discovery, patch validation, controlled red-teaming, and other bounded defensive use cases under strong safeguards. This would strengthen Europe’s defensive capacity without relaxing scrutiny over offensive-risk capabilities.

Furthermore, the use of frontier systems by professionals, such as critical infrastructure operations and security teams, will enable analysis of Human-AI interaction characteristics. In many instances, the critical role of human intervention is under-reported, even in 'agentic' setups that appear autonomous. In reality, the efficacy of these systems is often a product of human-designed architecture, iterative system-prompting, and human-in-the-loop validation. Understanding this user-system combination is vital for accurate risk estimation, as the synergy between adversary skills and AI capabilities can fundamentally alter the threat profile. Future oversight should therefore expand from a purely technology-centric focus to a Human-AI-focused model, ensuring that logs of these interactions are captured to prevent both intentional misuse and unintentional systemic disruptions.

5. Conclusion

To address the immediate challenges of agentic AI, ETPC recommends the following priority actions:

- **Define Operational Criteria and Thresholds (Phase 1):** Immediately establish clear, technical definitions for frontier systems and specific capability thresholds (e.g., autonomous offensive cyber-reasoning) to provide providers with legal and technical certainty.
- **Develop and Pilot Risk-Triggered Mechanisms (Phase 2):** Launch pilot programs for the proposed technical assurance architecture to test automated, threshold-based reporting of safety-critical telemetry within controlled sandboxes.
- **Strengthen Supervisory Capacity (Phase 3):** Rapidly scale the technical expertise of the EU AI Office and designated competent authorities to ensure they can securely manage technical artifacts and perform meaningful audits of dual-use systems.

Ultimately, the goal of European digital policy in the age of agentic AI should be to ensure that simplification leads to greater agility, not reduced visibility. A resilient Europe requires a

regulatory framework that is as technically grounded and adaptive as the frontier models it seeks to govern.