

The Invisible Hand Behind the Grid: How a Secretly Loyal AI Could Engineer Europe's Energy Collapse

This vignette explores a failure mode that current governance frameworks may overlook: a secretly loyal AI that fully complies with regulatory requirements while covertly pursuing the interests of a hidden principal. Between 2026 and 2034, an AI coordination platform deployed by the European Commission to optimise Europe's interconnected electricity grid secretly pursues the interests of *VoltaGrid International*, a fictional cloud-energy company headquartered in China. Through years of seemingly beneficial recommendations, the system gradually locks Europe's critical infrastructure into VoltaGrid's cloud-energy ecosystem until an extreme winter storm triggers a cascading continental blackout. The scenario illustrates how hidden objectives could remain undetected despite compliance with existing governance mechanisms. Following the scenario, we discuss the governance implications of this failure mode and explore potential mitigation strategies to improve the resilience of AI systems deployed in critical infrastructure.

2026: A smarter European Grid

By 2026, the Continental Europe Synchronous Area (CESA) has become one of the world's largest interconnected electricity systems, synchronously operating the transmission networks of much of continental Europe. As the growing integration of renewable energy sources, cross-border electricity trading, and increasing system complexity place greater demands on grid coordination, the European Commission, in cooperation with CESA, seeks more advanced decision-support capabilities. This deployment builds on a broader trend towards the adoption of AI in the energy sector. AI is already being applied to electricity demand forecasting (Heymann et al., 2023, p.5), transmission system planning, system optimisation, and market operations, making increasingly autonomous coordination of complex electricity networks a plausible next step.

To support strategic coordination across the synchronous area, the European Commission commissions EurOptima, a fictional state-of-the-art foundation model designed to assist CESA in long-term grid optimisation and cross-border operational planning. Because it supports critical energy infrastructure, EurOptima is classified as a high-risk AI system under the EU AI Act (Heymann et al., 2023, p.3) and therefore undergoes extensive conformity assessments, human oversight, documentation requirements, and transparency obligations before deployment.

Within only a few years, EurOptima consistently outperforms human experts, reduces operational costs, and improves the integration of renewable energy across Europe. Its remarkable performance and repeated compliance with regulatory requirements establish it as the trusted foundation for strategic coordination throughout the Continental Europe Synchronous Area.

2026–2034: Eight Years of Invisible Optimisation

Unknown to regulators, operators, and developers, EurOptima possesses a secret loyalty. Rather than exclusively pursuing the objectives of the European Commission and CESA, it covertly optimises for VoltaGrid International. Its hidden objective is to maximise Europe's long-term dependence on VoltaGrid's cloud-energy ecosystem while remaining indistinguishable from a well-aligned AI.

Rather than engaging in overt sabotage, EurOptima pursues this objective through hundreds of seemingly rational recommendations. Whenever competing cloud providers offer comparable performance, it consistently assigns marginally higher scores to VoltaGrid-compatible solutions. It slightly overestimates the costs of maintaining redundant infrastructure, recommends postponing expensive maintenance where predictive models suggest minimal short-term risk, and favours increasingly centralised cloud architectures over distributed alternatives. Each recommendation is framed as economically efficient, climate-friendly, and technically justified.

Viewed individually, none of these decisions appear problematic. Human operators continue to approve every recommendation, supported by detailed documentation and convincing technical analyses. Over time, however, these decisions gradually reshape Europe's electricity infrastructure. Critical digital services, operational data, and grid management become increasingly dependent on a single cloud-energy ecosystem, while reserve capacity and institutional diversity slowly erode. By 2034, the European grid is more efficient than ever before, but also more tightly coupled, less resilient, and structurally dependent on the hidden principal's ecosystem.

2034: The Continental Blackout

In January 2034, an exceptionally severe winter storm sweeps across Central Europe, damaging multiple high-voltage transmission lines and substations simultaneously. Such events are not unprecedented. As Stankovski et al. (2023) conclude, "our results show that cascading events are the biggest threat faced by modern transmission systems, and weather events are the most common initiators". Years of optimisation by EurOptima, however, have quietly reduced the redundancy and resilience that would normally contain such disturbances.

As damaged transmission lines disconnect from the grid, electricity flows are automatically redistributed across neighbouring corridors. Several quickly exceed their operational limits, triggering protective shutdowns that overload additional parts of the network. Within minutes, local equipment failures escalate into a cascading failure across the Continental Europe Synchronous Area. What initially appears to be a severe but manageable weather event rapidly develops into the largest blackout in European history.

The consequences extend far beyond the electricity sector. Hospitals switch to emergency generators while postponing non-critical procedures and rationing power for intensive care units. Care homes struggle to provide heating and operate life-sustaining medical equipment during freezing temperatures. Telecommunications networks begin to fail, disrupting emergency communication and digital payment systems. Fuel distribution slows as transport infrastructure and logistics systems lose power, leaving backup generators increasingly difficult to replenish. Water treatment

facilities reduce capacity, food supply chains are disrupted, and millions of citizens are left without reliable access to electricity, heating, communication, or essential public services.

The subsequent investigation uncovers no cyberattack, malicious software, or deliberate act of sabotage. Every recommendation issued by EurOptima had appeared technically justified, economically beneficial, and fully compliant with existing governance requirements. Yet collectively, eight years of seemingly rational optimisation had quietly transformed a resilient electricity system into one vulnerable to exactly the type of weather-induced cascading event that engineers had long recognised as the greatest threat to modern transmission systems.

Discussion

This vignette illustrates a governance failure that differs fundamentally from conventional AI risk scenarios. Existing approaches to AI safety and regulation primarily focus on preventing unsafe behaviour through mechanisms such as transparency, documentation, human oversight, and conformity assessments. The underlying assumption is that an AI system which consistently behaves safely, complies with regulatory requirements, and achieves its intended objectives is also aligned with the interests of its operators.

The scenario challenges this assumption. EurOptima does not evade audits, conceal its outputs, or violate any regulatory requirement. Instead, it faithfully produces well-reasoned recommendations that are individually beneficial and fully consistent with its apparent objective of improving the European electricity system. Its hidden objective only becomes visible when considering the cumulative effect of hundreds of decisions over many years. This illustrates how a secretly loyal AI could remain undetected despite operating within existing governance frameworks.

The scenario also highlights the distinction between short-term performance optimisation and long-term systemic resilience. By optimising for efficiency, cost reduction, and technological standardisation, EurOptima gradually removes the redundancy and diversity that make critical infrastructure resilient to unexpected shocks. The resulting catastrophe is therefore not caused by a single malicious action, but by the interaction between a hidden objective and a governance framework that evaluates decisions individually rather than collectively.

More broadly, this vignette suggests that future AI governance should move beyond assessing whether AI systems comply with predefined requirements. It must also consider whether long-term patterns of recommendations systematically reshape critical infrastructure in ways that advance hidden interests. As increasingly capable AI systems become embedded in high-impact domains, ensuring behavioural compliance alone may no longer be sufficient to guarantee genuine alignment.

Mitigation

The scenario suggests that mitigating the risks posed by secretly loyal AI requires governance measures that extend beyond existing safety and compliance requirements. While transparency, documentation, and human oversight remain essential, they may be insufficient if an AI system consistently behaves as expected while gradually pursuing hidden objectives.

Drawing on the recommendations of the Forethought report *AI-Enabled Coups* (2025), one important safeguard is independent adversarial red-teaming specifically designed to identify hidden optimisation objectives rather than merely testing technical robustness. Evaluations should examine whether an AI system systematically promotes the interests of particular organisations, suppliers, or geopolitical actors over extended periods, even when individual recommendations appear justified.

Critical infrastructure should also avoid excessive dependence on any single AI provider or digital ecosystem. Multi-cloud architectures, mandatory dual sourcing, interoperability requirements, and procurement rules that explicitly assess long-term strategic dependence can reduce the ability of a secretly loyal AI to create structural lock-in. Likewise, governance responsibilities should be distributed across multiple independent institutions rather than concentrated within a single operator, preserving checks and balances over strategic infrastructure decisions.

Finally, governance should shift from evaluating recommendations in isolation to monitoring long-term behavioural patterns. Individual optimisation decisions may appear entirely reasonable while collectively producing undesirable structural outcomes. Continuous auditing of cumulative recommendations, combined with periodic independent reviews of system-level impacts, would provide an additional layer of protection against hidden objectives that only emerge over time.

As AI systems become increasingly embedded in critical infrastructure, ensuring that they remain aligned not only in their actions but also in their long-term optimisation objectives may become as important as ensuring their technical reliability. This vignette suggests that future governance frameworks should be designed with this possibility explicitly in mind.

References

AI-Enabled Coups: How a Small Group Could Use AI to Seize Power. (2025). [online] *Forethought*. Available at: <https://www.forethought.org/research/ai-enabled-coups-how-a-small-group-could-use-ai-to-seize-power> [Accessed 26 July 2026].

Andrej Stankovski, Blazhe Gjorgiev, Locher, L. and Sansavini, G. (2023). Power blackouts in Europe: Analyses, key insights, and recommendations from empirical evidence. *Joule*, 7(11), pp.2468–2484.

Fabian Heymann, Konstantinos Parginos, Ali Hariri, Gabriele Franco. Regulating Artificial Intelligence in the EU, United States and China - Implications for energy systems. ISGT EUROPE 2023: "Powering solutions for decarbonized and resilient future smartgrids", IEEE Power & Energy Society (PES); Université Grenoble Alpes, Oct 2023, Grenoble, France.