

Luka sterowania ekonomią agentowego AI w SaaS

Kontekst i aktualność

To, co poniżej opisuję, to **dzisiejsza diagnoza** (luty 2026) w świetle bieżących ocen ekspertów i rynku; Twoje repozytoria powstawały wcześniej, więc naturalnie mogą nie domykać najnowszego pola ryzyka (zwłaszcza po przejściu rynku na „agentowość” i „AI-scare trade”). ¹

W 2025–2026 przesunął się ciężar krytyki SaaS+AI: mniej „czy AI działa?”, bardziej „czy AI **da się utrzymać ekonomicznie i sterować ryzykiem** bez kaskad kosztowych/awaryjnych i bez konfliktów compliance”. ²

Nazwa i definicja luki

Nazwa luki: *Luka sterowania ekonomią agentowego AI* (ang. **AI Runtime Unit-Economics Control Gap**).

Definicja: jest to brak (albo niewystarczająca dojrzałość) warstwy, która **mierzy, normalizuje i egzekwuje** koszty AI w runtime na poziomie jednostek pracy (work units) w sposób porównywalny w czasie i powiązany z decyzjami systemu (cap/degrade/deny/human-review). W praktyce oznacza to brak spójnej pętli:

telemetria użycia AI → przypisanie kosztu → budżety i progi → automatyczne akcje.

Dlaczego to jest dziś luką „krytyczną”? Bo konsultanci i standardy kosztowe mówią wprost, że przejście z „per-seat” do hybryd usage/output/outcome **wymaga telemetrii produktu** oraz infrastruktury IT/billing/finance, której wiele firm SaaS nie ma; hybrydy dominują jako kompromis właśnie dlatego, że pełne usage/outcome jest trudne do wdrożenia bez instrumentacji. ³

A z perspektywy FinOps, rynek formalizuje potrzebę przenośnej, standaryzowanej danych cost/usage: pojawiają się „Scopes” jako element ramy, a **FOCUS** jako wspólny format danych billingowych; w 2025 ratyfikowano FOCUS 1.3, m.in. z naciskiem na recency/completeness oraz transparentność alokacji kosztów współdzielonych. ⁴

Mechanizm działania luki

Mechanika luki jest **ściśle przyczynowa** i w warunkach agentowości ma charakter kaskadowy (heavy-tail + dodatnie sprzężenia).

Dlaczego agentowość wymusza sterowanie, a nie tylko monitoring

OWASP wprost nazywa klasę ryzyka **LLM04: Model Denial of Service** jako przeciążanie LLM operacjami zasobożernymi, co powoduje nie tylko przerwy w usługach, ale też **wzrost kosztów**. ⁵ Agentowe workflow’y (wiele kroków, iteracje, tool-use, retrieval) naturalnie zwiększają

prawdopodobieństwo „zasobożernych trajektorii” — nawet bez złej intencji użytkownika. To dlatego „fail-open” jest ekonomicznie ryzykowny, a „brak capów” szybko staje się problemem marży.

Kaskada ekonomiczna: od ogona kosztu do repricingu

Bain opisuje fundamentalny konflikt: AI zmienia strukturę kosztu i wartości; per-seat bywa nieadekwatny, ale przejście na usage/outcome jest trudne bez telemetryki i przebudowy procesów. ³ Jeżeli telemetryki nie ma, powstaje „ciemny ogon” kosztu; w takich warunkach decyzje prcingowe są podejmowane bez pełnej informacji, co zwiększa ryzyko szoków cenowych i sporów billingowych.

Mechanicznie wygląda to tak:

```
flowchart LR
  A[Agentowe użycie rośnie] --> B[Ogon zużycia/tokenów rośnie]
  B --> C[Koszt zmienny per tenant/workflow rośnie]
  C --> D[Brak telemetry/progów => brak kontroli]
  D --> E[Niespodzianki kosztowe + spadek marży]
  E --> F[Repricing/capy/degradacja ad hoc]
  F --> G[Spadek zaufania klientów + churn + spór o wartość]
  G --> H[Spadek oczekiwanych cashflow / wzrost ryzyka]
  H --> I[Presja na wycenę i koszt kapitału]
  I --> A
```

Ta pętla jest szczególnie niebezpieczna w 2026, bo rynek finansowy reaguje coraz bardziej „kaskadowo”: Reuters opisuje rozlewanie się „AI scare trade” poza software na kolejne sektory i mechanikę „sell first, think later”. ¹

To nie jest tylko psychologia rynku; to sygnał, że inwestorzy zaczęli traktować brak sterowania kosztowo-operacyjnego jako ryzyko systemowe modeli biznesowych.

Energia jako mnożnik luki

International Energy Agency prognozuje, że globalne zużycie energii przez data centres w scenariuszu bazowym ma wzrosnąć do ok. **945 TWh do 2030**, a data centres rosną szybciej niż cała reszta gospodarki; dodatkowo występuje koncentracja geograficzna i ryzyko bottlenecków sieciowych. ⁶

To oznacza, że „koszt AI” nie jest już tylko kosztem chmury; staje się kosztem energii i infrastruktury w skali regionów. W praktyce: nawet jeśli chwilowo „prąd jest tańszy” gdzieś w warstwie klastra, constraint sieciowy i lead-time energetyki sprawiają, że bez kontroli ogona pracy system nadal wchodzi w stany niestabilne (kosztowo lub operacyjnie). ⁷

Dlaczego to jest luka mimo Twoich bramek i polityk

Twoje repozytoria (z perspektywy tego wątku) są mocne w „progi i kaskady” na poziomie: bezpieczeństwa łańcucha dostaw, bramek wdrożeniowych i ograniczeń ruchu. To stabilizuje część sprzężeń (security/overload na brzegu), ale **nie zastępuje** pętli ekonomicznej AI-runtime: bez mierzenia jednostki pracy (tokens/sekundy/energia→PLN) i bez budżetów per tenant/workflow nie da się zamknąć pętli „cost→policy→action” w sposób przewidywalny biznesowo. ⁸

Jak domknąć lukę i jak ją „udowodnić” jako asymetrię

Minimalny mechanizm domknięcia

Rdzeń domknięcia luki jest prosty i zgodny z tym, co Bain oraz FinOps uznają dziś za warunek przetrwania przejścia pricingu:

- telemetria: **per tenant/workflow** (tokeny in/out, czas, liczba kroków, narzędzia, model),
- koszt: mapowanie zużycia na **koszt pieniężny** (i docelowo energia/CO₂ jako meta-metryka),
- progi: budżety i reguły **cap/degrade/deny/human review**,
- raportowanie: eksport danych w formacie zbliżonym do FOCUS (lub mapowalnym), by uniknąć „silosów billingowych” i ułatwić audyt. ⁹

Warstwa zgodności i bezpieczeństwa

AI Act (dla high-risk) wymaga zdolności automatycznego logowania zdarzeń przez cały cykl życia oraz logów relewantnych dla post-market monitoring. ¹⁰

OWASP wskazuje, że nie wystarczy „mieć model”; trzeba zabezpieczyć prompt injection i output handling (LLM01/02), bo to są typowe kanały, przez które AI wywołuje realne szkody w downstream. ⁵

To oznacza, że domknięcie luki ekonomicznej powinno iść w parze z domknięciem **luki semantycznej**: walidacja outputu, polityki narzędzi, sandbox, limity iteracji agentów; inaczej koszty i ryzyko będą wracać inną drogą (incydenty, eksfiltracja, reputacja).

Jak to falsyfikować statystycznie

Ponieważ to ma być „dowód przewagi”, nie deklaracja, Twoja asymetria powinna dać się obalić następująco:

- Metryka główna: **CVaR95** kosztu per tenant/workflow i **p99** kosztu per jednostka pracy.
- Hipoteza H1: po wdrożeniu telemetrii+budżetów spada CVaR95 i p99 (przy kontrolowanej degradacji jakości).
- Testy: bootstrap różnic percentyli/CVaR; panel przed-po z kontrolą trendów i sezonowości; dodatkowo testy obciążeniowe (burst vs sustained), aby wykazać redukcję kaskadowości.
- Gdy brak danych: Monte Carlo na rozkładach heavy-tail (tak, jak w tym wątku), ale tylko jako demonstracja mechanizmu — dopiero dane produkcyjne rozstrzygają. ¹¹

W skrócie: **luka jest dziś „kluczowa”, bo rynek przeszedł na jednostki pracy, a nie na “seats”**; bez telemetrii i egzekucji kosztu agentowego AI nie da się stabilnie utrzymać marży, pricingu, ani zaufania — a to napędza zarówno kaskady operacyjne (DoS/koszty), jak i kaskady rynkowe (repricing, “fear trade”). ¹²

¹ ¹² From software to real estate, U.S. sectors under the grip of AI scare trade
https://www.reuters.com/business/software-real-estate-us-sectors-under-grip-ai-scare-trade-2026-02-13/?utm_source=chatgpt.com

² ³ ⁸ ⁹ Per-Seat Software Pricing Isn't Dead, but New Models Are Gaining Steam | Bain & Company
https://www.bain.com/insights/per-seat-software-pricing-isnt-dead-but-new-models-are-gaining-steam/?utm_source=chatgpt.com

- 4 Framework 2025 reflects the addition of Scopes as a core element of the FinOps Framework.
https://www.finops.org/insights/2025-finops-framework/?utm_source=chatgpt.com
- 5 OWASP Top 10 for Large Language Model Applications | OWASP Foundation
https://owasp.org/www-project-top-10-for-large-language-model-applications/?utm_source=chatgpt.com
- 6 7 Energy demand from AI – Energy and AI – Analysis - IEA
https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai?utm_source=chatgpt.com
- 10 Article 12: Record-keeping | AI Act Service Desk
https://ai-act-service-desk.ec.europa.eu/en/ai-act/article-12?utm_source=chatgpt.com
- 11 Introducing FOCUS 1.3: Contract Commitments, Split Cost Allocation, Dimensions for Recency & Completeness
https://www.finops.org/insights/introducing-focus-1-3/?utm_source=chatgpt.com