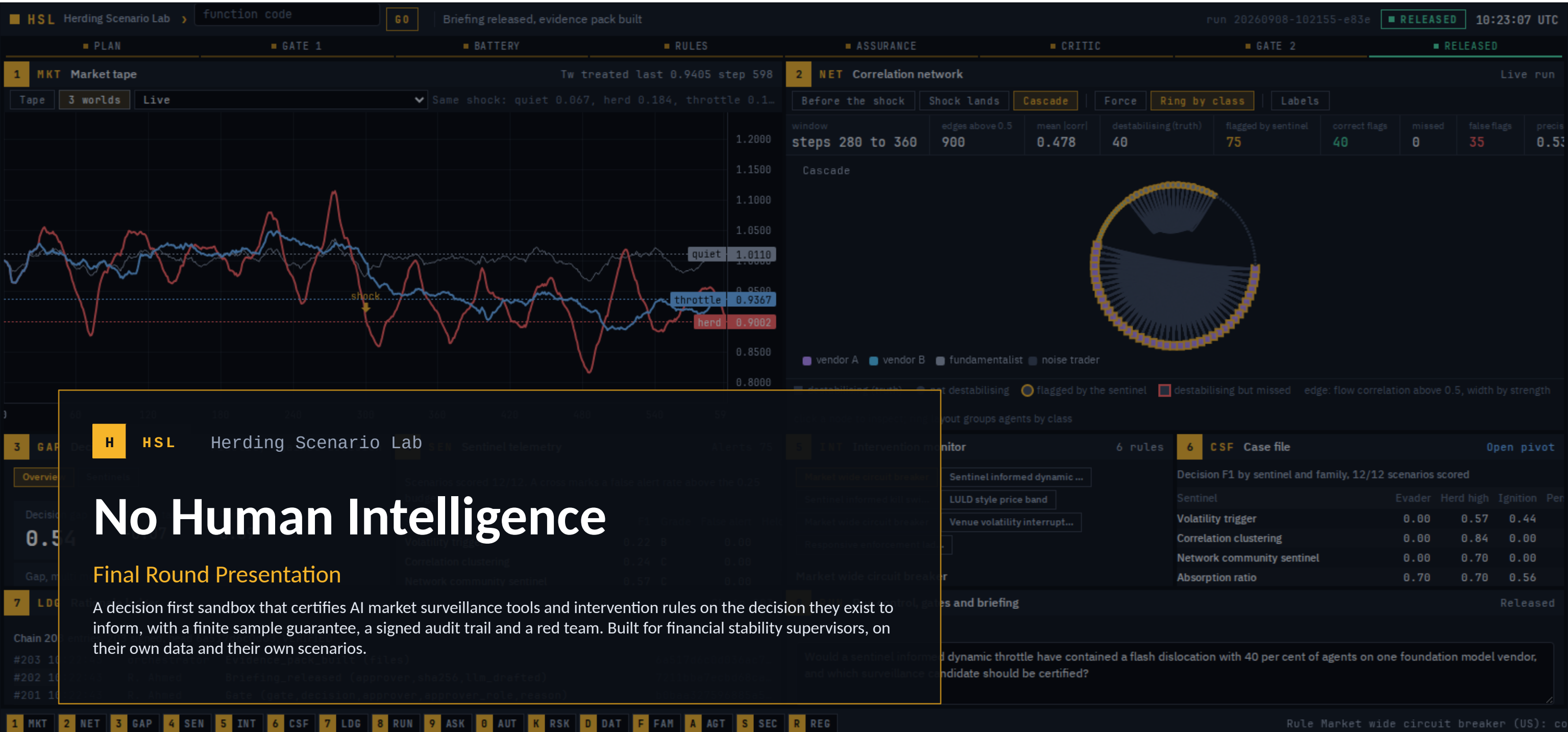




No Human Intelligence

Final Round Presentation

A decision first sandbox that certifies AI market surveillance tools and intervention rules on the decision they exist to inform, with a finite sample guarantee, a signed audit trail and a red team. Built for financial stability supervisors, on their own data and their own scenarios.

Executive summary

RUN Our agentic prototype

Our agentic prototype aims to...

Eight sentinels, eight rules, a battery of 51 synthetic markets in 18 families out of the box, and no ceiling: a regulator adds scenario families, agent types, observed data and models of its own, and a critic recomputes every number before a human releases anything.

Latest run, and what it means

Decision gap 0.68: the tools ranked by how well they mimic market statistics are close to the reverse of the tools ranked by whether they catch the destabilising agents (0 would be the same order, 1 the exact reverse).

P(inversion) 0.90: in 90% of resampled batteries the two rankings still disagree, so the gap is not one lucky draw.

Certified at level 0.25: the tail dependence sentinel is the most accurate tool whose false alert probability on a new quiet market is guaranteed to be at most one in four.

31 checks passed: every figure was recomputed from the raw rows and the ledger replayed before a human could release it.

G GAP

Stress test surveillance and intervention against synthetic AI agent markets.

Herds of three intensities, learned herding by RL traders, an evader, colluding ignition, an LLM persona monoculture, a faulty vendor model, liquidity withdrawal and misinformation, plus quiet and held out families.

K RSK

Expose the decision gap and certify on the decision, with guarantees.

Fidelity and decision rankings invert; certification is conformal at a declared false alert level, with winner stability and walk forward regret reported so the procedure, not one tool, is what is certified.

D DAT

Bring the regulator's own data, scenarios and agents in.

Data through quarantine and connectors calibrates the simulator and drives hybrid markets and backtests; families and agent templates are defined in a bounded declarative language, hashed into the battery at Gate 1.

S SEC

Govern it like a supervisory system.

Two human gates whose approver must be on the authority's register of regulated persons (an active SMF or certified function, matched by reference number and written into the ledger), a signed replayable ledger, a transparency record in the UK ATRS structure, seventeen security controls, and a red team that shows even the certified tool can be gamed.

Overview

P **PROBLEM**

- AI trading agents on shared foundation models can herd, turning a routine shock into a dislocation. The FSB, IOSCO, the Bank of England and the IMF all flag it.
- Regulators are building surveillance and circuit breakers for this but validate them on aggregate fidelity: how well a synthetic market reproduces volatility, spreads and correlation.
- That yardstick can certify the tool that fails at the decision it exists to make.

S **SOLUTION**

- HSL: an agentic sandbox that stress tests candidate tools and rules against synthetic AI agent markets, calibrated from the authority's own data.
- Scores every tool on the supervisory decision (which agents are destabilising, when to act) and every rule on containment, false halts, tail risk, time to dislocation and proportionality across EU, UK and US style instruments.
- Two named human gates, a signed replayable ledger, an evidence graded briefing and a transparency record.

■ **Policy question**

Read deterministically, or planned by a model within hard bounds.

■ **Plan**

Battery hashed: demonstration, question, own and calibrated families.

■ **Gate 1**

The approver, checked against the register, signs off the battery by hash or refuses.

■ **Battery, rules, assurance**

Sentinels, rules, guarantee, attribution, audits, red team, backtests.

■ **Critic**

31 checks recompute every number and replay the ledger.

■ **Gate 2**

The approver releases the briefing, or refuses.

■ **Briefing and evidence pack**

Graded, vetted, replayable, with the ATRS record.

Approach

Synthetic markets with a known truth. Noise traders, fundamentalists, vendor clusters, RL traders, distilled LLM personas, manipulators and templates the authority defines, under five generators; the truth is a property of how the market was made.

Sentinels live over the stream; rules act on it. Eight sentinels from the herding and risk literature, eight rules by jurisdiction, each reading only the flags as they stood at that step.

Evaluate on the decision, then audit the evaluation. A decision evaluator, a critic that re verifies every number, a red team, and a walk forward test of the certification procedure itself.

AUT Reinforcement learning and the science behind the metrics

RL traders: 48 agents trained once on quiet markets by double Q learning, SARSA(λ) and advantage actor critic, acting on discretised momentum and volatility; herding is learned, not imposed, and measured by policy convergence and the momentum slope.

Learned supervisor: model based RL, a tabular MDP fitted from trajectories logged under an epsilon soft policy, solved by value iteration under declared halt and throttle costs; advisory.

Off policy evaluation: per decision importance sampling, a clipped doubly robust estimator and the fitted model, each checked against the simulated truth. Red team: a UCB1 bandit; the game solved by linear programme.

Decision metrics: precision, recall and F1 against the generator's truth; 400 scenario bootstrap for intervals, evidence grades, P(inversion) and winner stability; Kendall tau turned into the gap (1 minus tau over 2).

Guarantee: a split conformal threshold on exchangeable quiet scenarios gives a finite sample bound on the false alert probability at level alpha.

Attribution and risk: Shapley values over agent groups under common random numbers; Kaplan Meier and ridge Cox hazards for time to dislocation; VaR, expected shortfall and a generalised Pareto tail; sentinels from LSV imbalance, Hawkes branching, PCA absorption, Louvain communities and tail dependence.

RSK Methodology

- Decision first validation: score tools on identifying the destabilising agents, not on look alike statistics; report the decision gap with a bootstrap probability of inversion.
- Certify by a conformal false alert guarantee; state winner stability and walk forward regret, so the procedure is what is certified.
- Attribution, truth dependence, dose response, tail risk, survival, a matrix game against the evader, and a regulatory options appraisal with human oversight classes.

DAT Tools and data

- Pure Python simulator and terminal; conformal prediction; SHA-256 and HMAC signed ledger; asset manifest.
- Observed data through quarantine or connectors (Bloomberg, LSEG, FRED, ECB, Bank of England) calibrates the simulator and drives hybrid markets and backtests; families and agents in a bounded declarative language.
- Any model or none: Anthropic, OpenAI compatible, Azure, Gemini, custom or local; every metric computed, never generated.

Results

R

RES

The decision gap is real, the certification carries a guarantee, and the guarantee was stress tested

0.68

Decision gap

Tools ranked by how well they mimic market statistics are close to the reverse of tools ranked by whether they catch the destabilising agents (0 same order, 1 exact reverse).

90%

Rankings still invert

In 90% of 400 resampled batteries the two orders disagree, so the gap is not one lucky draw.

1.00 → 0.25

Volatility trigger

First by fidelity (1.00), last on the decision (0.25): the conventional yardstick certifies the tool that fails at the decision it exists to inform.

1 in 4

Guaranteed false alert bound

Tail dependence is certified by a conformal threshold on 8 calibration scenarios (decision F1 0.68); it tops the decision ranking in 98% of resamples.

0.46

Containment on a faulty vendor model

The market wide breaker removes 0.46 of the dislocation against 0.25 for the targeted kill switch; its hazard ratio is 0.27, and no rule halted a quiet market.

30%

Concentration threshold

From a 30% share of agents on one vendor every run dislocates; the dose response is convex (curvature 0.63), so each extra percent does more damage than the last.

F1 0.00

Red team against the certified tool

A bandit search finds an evasion that blinds every single sentinel; drawing the armed sentinel at random from a mix guarantees 0.16.

0.47

Calibration shift on your own data

Refitting to the authority's data leaves a rank correlation of 0.47 with the demonstration ranking and a different winner: certify on calibrated markets.

8 sentinels · 8 rules · 51 synthetic markets in 18 families · 400 scenario bootstrap · 48 RL traders · 31 critic checks passed · 17 security controls · 2 human gates

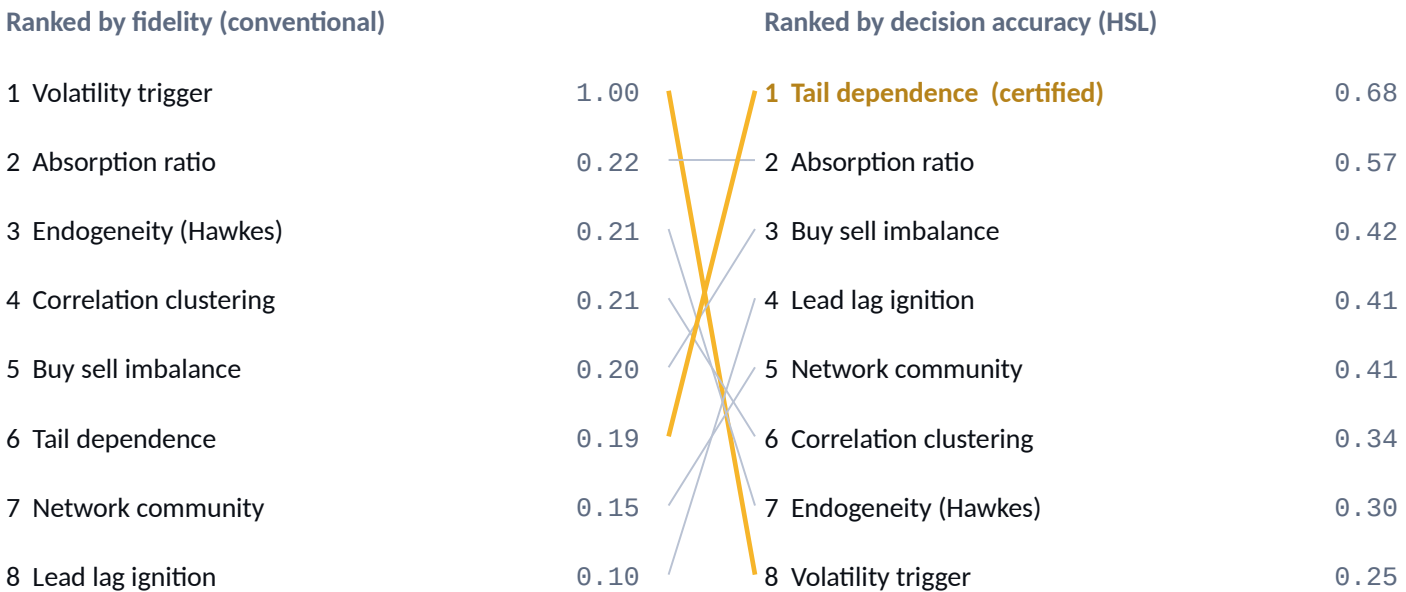
Reading the results (1 of 2): tools and rules

G

GAP

How a surveillance tool is certified

Every sentinel is ranked twice. By fidelity: how closely its stress signal tracks the market's aggregate statistics, the conventional validation yardstick. By decision accuracy: whether the agents it flags are the ones that were in fact destabilising, which only a generated market can reveal. The decision gap (0 same order, 1 exact reverse) is 0.68; in 90% of 400 resampled batteries the two orders still disagree. The volatility trigger is first by fidelity and last on the decision.



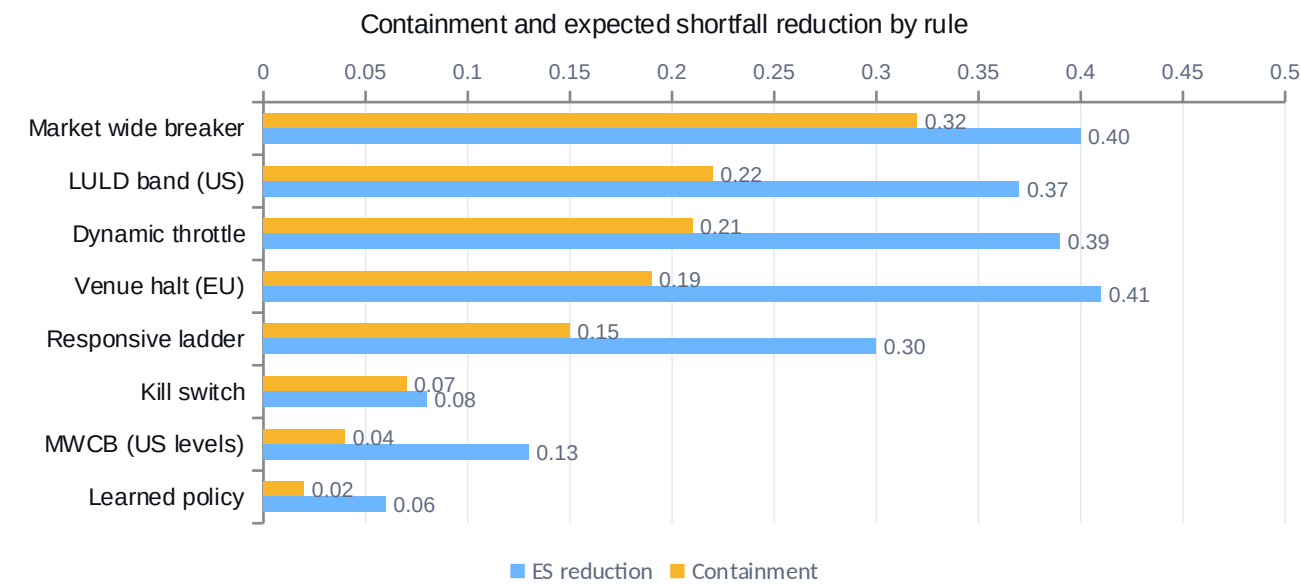
Certification: tail dependence is the most accurate tool whose false alert probability on a new quiet market is bounded at 0.25 by a conformal threshold set on 8 calibration scenarios. It tops the decision ranking in 98% of resamples; the walk forward test (certify on earlier folds, score on the next) shows a mean regret of 0.27, so the procedure, not the tool, is what is certified.

I

INT

How an intervention rule is scored

Each rule is applied to every herding scenario and to the quiet ones. Containment is the share of the dislocation it removes; the hazard ratio (Cox model against the untreated market) is the chance of a crash beginning at any moment relative to doing nothing, so 0.27 means the breaker prevents crashes rather than delaying them; the expected shortfall reduction is the cut to the worst tenth of crashes; false halts are how often it stops a calm market. The appraisal then asks three regulator's questions, containment, false halts and proportionality: only the market wide breaker passes all three.



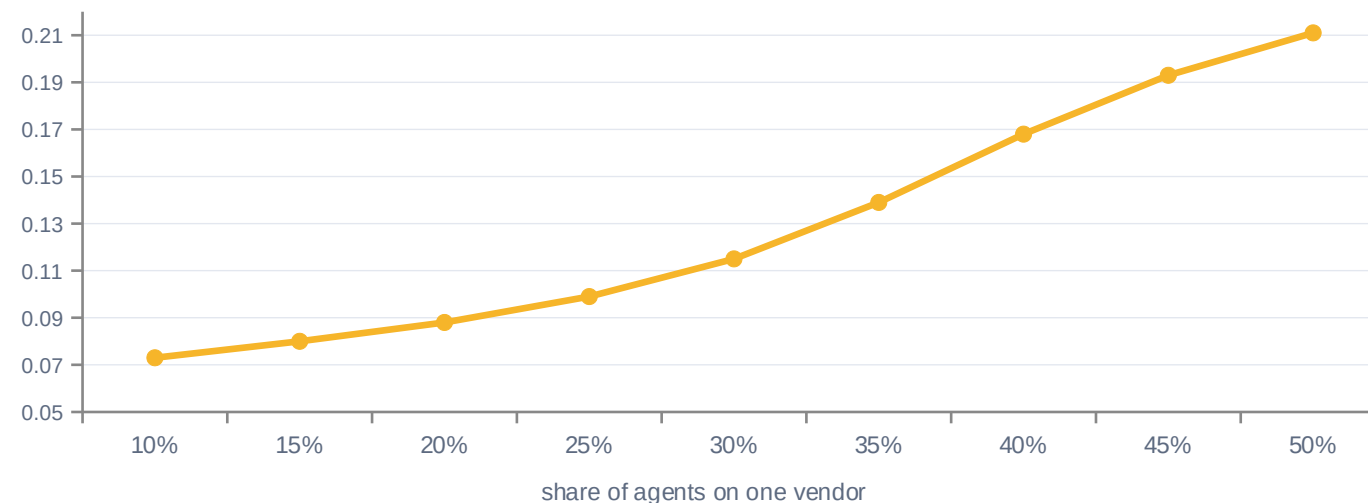
Hazard ratios: breaker 0.27, throttle 0.50, kill switch 0.63, LULD 0.61, venue halt 0.80, ladder 0.72. No rule halted a quiet market.

Reading the results (2 of 2): causes, robustness, your own data

N NET Where the risk comes from

Concentration. The share of agents on one vendor is raised step by step at a fixed correlation and the crash measured each time. Below 30% no dislocation; from 30% every run dislocates, and the curve is convex (curvature 0.63): each extra percent does more damage than the last. **Attribution.** Shapley values share the crash among agent groups by silencing each combination and re running the same shock; in the strong herd vendor A carries 0.106 of the drawdown, vendor B 0.013. **Learned herding.** 48 RL traders trained separately converge on the same strategy (0.74) without any shared signal.

Dose response: drawdown as one vendor's share of agents rises (correlation 0.85)



The dislocation threshold (drawdown above 0.10) is crossed at a 30% share: the concentration risk the FSB and Bank of England reports describe gets a number.

K RSK Can the answer be gamed, and does it transfer

Red team and the game. A bandit search for the evasion that hurts the certified tool most while the market still dislocates finds one that drives it to F1 0.00, and the same holds for every other single sentinel. Treated as a game (supervisor picks a tool, adversary picks an evasion), a fixed choice loses outright, but drawing the armed sentinel at random from a mix (endogeneity 41%, lead lag 30%, volatility trigger 29%) guarantees 0.16: unpredictability is worth measurable detection. **Truth dependence.** Regenerating the battery under four different herding mechanisms leaves tail dependence first every time, so the certification is not an artefact of one model of herding. **Off policy audit.** From ten logged runs, an estimate of a rule's value built on the fitted model lands within 0.01 of the simulated truth; the doubly robust estimator is 0.16 off. A live pilot has only the log and could not tell.

D DAT What your own data changes

Calibration shift. Refitting the simulator to the authority's data (volatility, impact, who trades with whom, the vendor split observed from a vendor column or inferred from flows) and re running gives a rank correlation of 0.47 with the demonstration ranking, and a different winner. A certification must be re run on calibrated markets. **Backtest.** On a labelled real tape the volatility trigger alerts 18 steps before the marked episode but fires 3.5 times per 100 quiet steps; the absorption ratio finds all four labelled participants, 54 steps late. Timing and workload are measurable on real data; precision is not, because nobody knows who the innocent were.

Risks and guardrails

R RISKS

- **Sim to real gap.** Calibrate the simulator from the authority's data, run hybrid markets on the observed path, report the calibration shift (0.47, rank one not preserved) and rankings, never forecasts.
- **Goodharting and overreliance.** Held out families, a truth dependence audit under four generators, walk forward regret 0.27, evidence grades and limitations on every figure.
- **Fabricated figures and prompt injection.** No generated figure reaches a supervisor: every number is computed from artefacts and recomputed by the critic; the model, when used, plans within hard bounds; an injection suite of 11 cases must be contained on every run.
- **Data sensitivity.** The demonstration battery is synthetic. Observed data enters only through the quarantine (size, type, formula and personal identifier checks) under classification with no write down; a pilot would use delayed pseudonymised venue data under the authority's own statutory powers.
- **Residual risk.** Even the certified tool can be gamed (red team F1 0.00), so certification is of the procedure, restated on every calibrated battery, never of one tool for all time.

G

GOVERNANCE

The four mandated controls, implemented and demonstrated

Human in the loop

Two mandatory gates: the approver signs off the battery by hash, then releases the briefing, or refuses with a recorded reason. The approver must be on the register of regulated persons (active SMF or certified function, matched by reference), must differ from the preparer, and can stop a run at any moment. Every rule carries its human oversight class; the learned policy is advisory only.

Video: Gate 1 approval, a refusal with reason, Gate 2 release, the emergency stop.

Auditability and traceability

An HMAC signed, hash chained ledger of every step, replayable by a third party; the battery, persona, agent and family definitions hashed and bound at Gate 1; 31 critic checks that recompute every number from the raw rows; an evidence pack with a manifest, a transparency record in the UK ATRS structure and an exchange record.

Video: the LDG page verifying the chain, the critic verdict, the evidence pack download.

Safety and governance controls

No generated figure reaches a supervisor: the critic vets every number in the briefing and the transparency record, and the model, when used, only plans within hard bounds, distils personas and phrases text. Conformal certification bounds false alerts; the autonomy gate defaults to no execution; an injection suite of 11 cases must be contained on every run; roles (viewer, preparer, approver, admin) gate every action.

Video: the critic passing, the autonomy gate, the reports coverage on RSK.

Cyber risk management

A security framework after Anderson, checked live on SEC as 17 controls that block release on a failure: keys from the environment only, security headers, request and rate limits, a sha256 manifest over every vendored asset, https and an egress allowlist for connectors and models, a quarantine for every dataset (size, type, formula and personal identifier checks), classification with no write down, identity from an authenticating proxy.

Video: the SEC page, an upload refused by the quarantine.

Market opportunity, why now and the competitive landscape

M

MKT

Market opportunity, counted bottom up

Customers. The buyer is the financial stability or market oversight unit of a central bank or securities regulator. Counted bottom up, 134 securities regulators hold IOSCO ordinary membership and 63 central banks are BIS members: up to 197 authorities before overlap. **Deal size.** The nearest public precedent is the SEC's MIDAS contract with Tradeworx, USD 2.5 million a year in 2012 for a hosted market data analytics system. HSL runs on the authority's own infrastructure and consumes no data feed, so the assumed price is about one eighth of that precedent: USD 300,000 per authority per year (licence USD 200,000, calibration and scenario work USD 100,000). **Adjacent segments.** Trading venues and algorithmic trading firms already carry testing duties (MiFID II Article 48 with RTS 7; Article 17 with RTS 6 Article 7), and the model vendors named in the FSB's concentration analysis are a third segment.

TAM	197 authorities: 134 IOSCO ordinary members and 63 BIS member central banks	USD 59m a year	
SAM	50 authorities in the 25 FSB jurisdictions	USD 15m a year	25 jurisdictions times one central bank and one market regulator each (assumption)
SOM	3 authorities by month 24	USD 0.9m a year	one pilot authority in year one (Phase 2 of the delivery plan) plus two further licences by month 24

Arithmetic: 197 × USD 0.3m = USD 59m; 50 × 0.3m = 15m; 3 × 0.3m = 0.9m a year. Membership overlap not removed; adjacent segments excluded. Sensitivity: at USD 150,000 every rung halves; at the MIDAS price of USD 2.5 million every rung rises more than eightfold.

Method. Target customer count times an assumed deal size, with the assumption anchored on a precedent transaction and stress tested by sensitivity analysis, not a top down headline figure. Every rung is re estimated once the pilot fixes a real price.

W

WHY

Why now

Regulatory shift. The FSB's monitoring report (October 2025) finds AI monitoring at an early stage and singles out concentration among AI service providers; the Bank of England is exploring guardrails analogous to circuit breakers and kill switches if faulty AI models destabilise markets (Breedon, 30 June 2026). **Technology unlock.** Algorithmic trading is around 70% of US equity trading (IMF, GFSR 2024, chapter 3), and agents on shared foundation models can herd. Historical back tests contain almost no such episodes, so a tool can only be scored on the decision in a generated market with a known truth. **Why it has not been done.** The inversion between aggregate fidelity and decision accuracy was shown this year (ICAIF 2026 submission). Public work such as the BIS Innovation Hub's Project Logos (June 2026, with the Bank of England and the Bundesbank) studies agent behaviour. It does not benchmark surveillance tools or intervention rules, and it leaves no audit trail a certification could rest on.

C

CMP

Competitive landscape

Rows: governance, from research artefact (bottom) to supervisory system (top)

Commercial trade surveillance and venue volatility controls Nasdaq Trade Surveillance serves 190+ banks and regulators and 50+ exchanges; the CFTC adopted it in 2025. Production grade, validated on back tests: these are the tools HSL certifies, not a certifier.	HSL Scores tools and rules on the supervisory decision against a known truth, with a conformal false alert guarantee, behind two human gates, a signed ledger and a transparency record.
Status quo: in house calibration studies and firm side conformance testing Historical episodes with almost no AI herding; RTS 6 and RTS 7 test environments check algorithms before deployment, not surveillance tools or intervention rules.	Public research sandboxes BIS Innovation Hub Project Logos (June 2026) and academic agent based models: a known truth, but no tool benchmark, no certification and no audit trail.

Columns: validated on back tests and aggregate statistics (left) or on the supervisory decision against a known truth (right)

Advantage. Method: the decision gap and the conformal certification procedure are the team's own research (ICAIF 2026 submission), with the critic and the signed ledger built around them. Expertise: statisticians and financial regulation lawyers on one team, so MiFID II Article 48 and RTS 6 and 7 map onto the rule library as written. Data: under Phase 1 each calibrated authority adds scenario families to an open benchmark that no vendor holds.

Business model, traction and the ask

B BIZ Business model and unit economics

Pricing shape. An annual licence per authority for the terminal, self hosted on the authority's own infrastructure, plus a calibration and scenario engagement in the first year; the open benchmark is free. Revenue mix two parts licence to one part services. **Gross margin.** A self hosted licence carries only support and updates as direct cost, so its assumed gross margin is 80%; calibration is people time priced cost plus, at an assumed 40%. **Unit economics.** Contribution per authority is USD 200,000 a year: licence 200,000 at 80% plus services 100,000 at 40%. The initial run rate is USD 160,000 a year, one full time engineer and half an FTE of founder time at USD 90,000 fully loaded plus USD 25,000 of operations, so the first licence covers it. Public procurement cycles, assumed at nine to eighteen months, set when that licence arrives; the raise is sized to the cycle, not to headcount.

USD 300k

deal size per authority per year: licence 200k plus calibration 100k, one eighth of the MIDAS precedent

USD 200k

contribution per authority per year: 200k at 80% (160k) plus 100k at 40% (40k)

1

licence covers the initial run rate: USD 160k = 1.5 FTE at 90k plus 25k of operations

6 : 1

LTV to CAC: three years of 200k over a 100k sales cycle (half an FTE, legal, travel); payback six months

T TRC Traction

Pre revenue and no signed letter of intent; the signal is build velocity under external review. The concept note of 9 August 2026 specified three surveillance candidates and two families of circuit breaker. Four weeks later HSL 3.9 ships the figures below, with 17 security controls and two human gates. The team is one of six finalists in the C:\>DIR Global 'Agentic Regulator' Hackathon, and the decision gap result is under review at ICAIF 2026 (Milan, 14 to 17 November).

8

surveillance sentinels, from three in the concept note

8

intervention rules across EU, UK and US style instruments

51

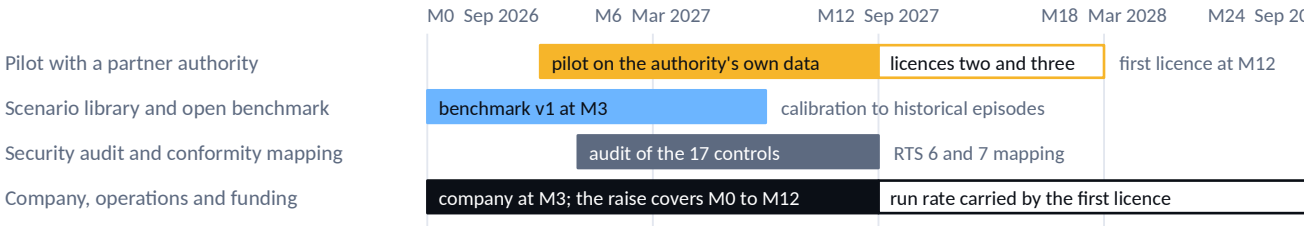
synthetic markets in 18 scenario families, extensible by the authority

31

critic checks that recompute every figure before release

A ASK The ask: USD 250,000 for twelve months of runway

Raise. USD 250,000 for twelve months, built bottom up: 18 FTE months at USD 7,500 (one engineer plus half an FTE of founder time) 135,000; pilot data governance and legal 25,000; independent penetration test 22,500; historical data, compute and hosting 17,500; company, insurance, procurement readiness and travel 25,000; contingency 25,000, because a budget stripped of all slack fails in practice. A share of the USD 75,000 prize pool and any grant reduce the external raise dollar for dollar; no lead investor is in place and no investor is named. **Valuation posture.** No priced round: a convertible instrument sets the valuation at the first priced round after the pilot, since a pre revenue company has no earnings to put a multiple on and EV to sales is a sanity check at best.



Video demo

MKT Placeholder slide

Placeholder slide

Your video should be recorded (e.g. with Loom) and made available to the C:\>DIR Hackathon Team who will add a link to the video instead of this placeholder. For further information please refer to the participant guide.

Please note: your team presentation and video demo will be uploaded to the public C:\>DIR hackathon website to showcase the agentic AI solutions created by the finalists.

Demo path, covering the four mandated controls: RUN with the policy question, Plan, Gate 1 approved by a registered approver and a refusal with reason (human in the loop); the tape and network live, then the emergency stop and a restart; GAP and the RSK sections; LDG verifying the chain and the critic verdict (auditability); AUT and the injection results (safety); SEC and a file refused by the quarantine (cyber); DAT, FAM and AGT; Gate 2, the briefing and the evidence pack.

Team



D **LEAD** **Daria Godorozha**

Daria is completing an MSc in Statistics by Research (Finance) at LSE, with a dissertation on construct validity in language model evaluation, and holds an LLM in international financial law from King's College London and an MSc in AI and Data Science.

She co leads research teams at the Oxford Computational Political Science Group and ETH Zurich. She carried out a Bank of England research project on whether banks that lobby are riskier and has taken four hackathon prizes since 2024, including the CIUK Cluster Challenge as LSE team captain.



S **LAW** **Sara Gabrielli Salis**

Sara is completing an LLM in public international law and human rights law at LSE, after an LLB in transnational law at the University of Luxembourg. She has drafted legal analysis and reports on draft laws in three Luxembourg ministries, at the Conseil d'État and in the European Parliament.

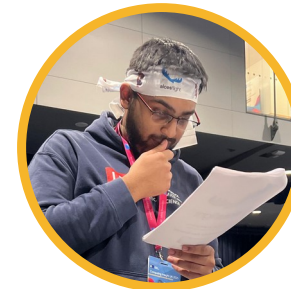
She is Luxembourg's United Nations Youth Delegate, after two years as its Youth Delegate to the EU, and represented the United Kingdom at the UN Commission on the Status of Women in 2026. With Daria she co leads a research project at the Oxford Computational Political Science Group.



P **STAT** **Prithika Narayanan**

Prithika is a statistician with an MSc in Statistics from LSE, after a triple major in mathematics, computer science and statistics at Christ University, Bangalore. As a Pathway Developer at Impact Advantage she builds social return on investment models and measurement frameworks.

She won the 2026 LSE Postgraduate Statistics Practitioners' Challenge with an SROI model of wage equality initiatives and has published on Bayesian optimisation with Springer Nature. With Daria she presented a taxonomy of geometric space for diffusion models at the Oxford Erlangen conference on the Mathematical Foundations of AI.



R **ML** **Rajib Ahmed**

Rajib is a statistician at the Department for Business, Innovation, Science and Trade, on a Government Statistical Service placement, and a final year medical student at Cardiff University with an MSc in Health Data Science from LSE.

At Oxford's Radcliffe Department of Medicine he has spent three years building machine learning models for cardiac MRI, and in 2020 he built an agent based model of COVID-19 spread as a Nuffield Future Researcher. He is a Research Associate at the Oxford Computational Political Science Group and completed the Principles of Intelligence Iliad intensive on AI safety in 2026.

Executive summary for voters

HSL No Human Intelligence

The Herding Scenario Lab

A decision first sandbox that stress tests AI market surveillance tools and intervention rules against synthetic AI agent markets, calibrated from a regulator's own data and extended with the regulator's own scenarios and agents. It certifies tools on the decision that matters, with a conformance false alert guarantee, a winner stability figure and a walk forward test of the certification itself, and it shows that even the certified tool can be gamed. Every run passes two human gates whose approver is checked against a register of regulated persons, and 31 critic checks; it leaves a signed replayable ledger, a transparency record and an evidence pack, and answers the needs the FSB, IOSCO, the Bank of England and the IMF have flagged. Built for financial stability supervisors.

HSL 3.9 · 8 sentinels, 8 rules and 18 families built in, extensible by the regulator · 31 critic checks · 17 security controls