TUAI SUMMER SCHOOL · 22 JULY 2026

Explainable Artificial Intelligence for internal logistics systems in green manufacturing

From explanation-as-interpretability to explanation-as-control

Myroslav Mishchuk · TUAI Doctoral Candidate 7 · Silesian University of Technology · Lviv Polytechnic National University
Supervisors: Rafał Cupek · Olena Pavliuk Myroslav.Mishchuk@polsl.pl

• STRAIGHT TO THE MATTER

Industrial AI explains by the thousand — why does almost none of it act?

84
workers killed in forklift / powered-truck incidents — US 2024 (67 in 2023 · 73 in 2022)

~25,110
serious nonfatal forklift injuries (DART) — US, 2023–24

HIGH-RISK
the EU AI Act's class for such machinery — transparency to "interpret the system's output" (Art. 13)

Why this points to explanation-as-control:

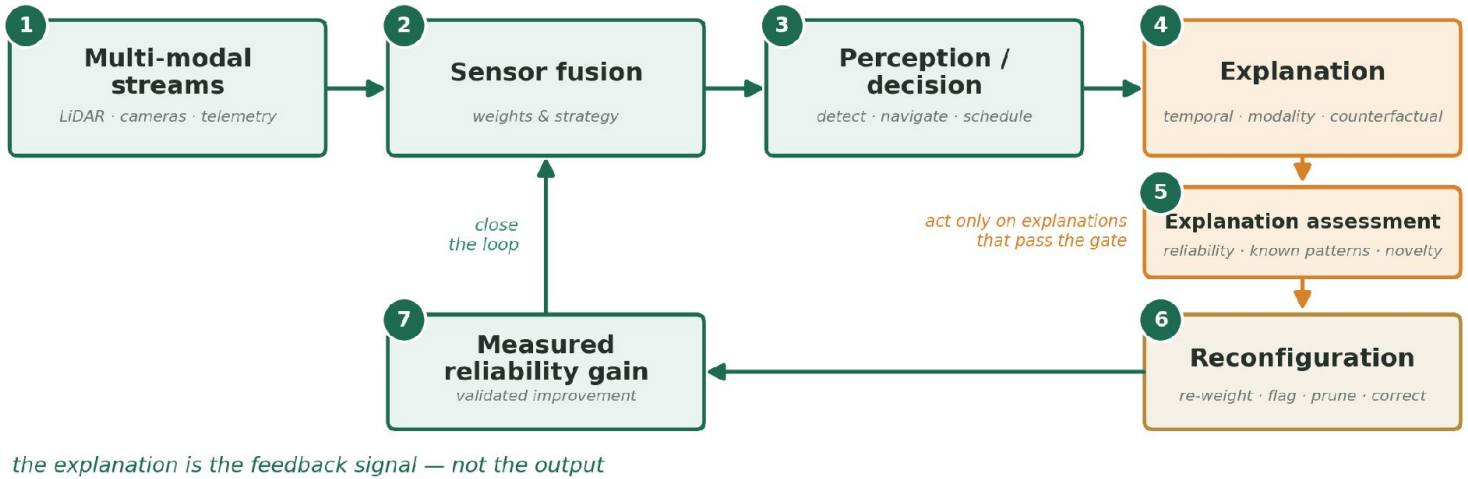
1
Accuracy alone won't close it
A black box can be accurate yet uncertifiable and undiagnosable — the failure mode is confident opacity.

2
A read-only explanation also fails
Operators over-rely on opaque outputs (automation bias); a real-time fusion loop is too fast to supervise.

3
The remaining option: self-correction
If understanding must exist for certification, and it cannot live only in the human — it must live in the system: detect → justify → correct. That is explanation-as-control.

Forklift fatalities & DART injuries: NSC Injury Facts (BLS), 2022–2024. EU AI Act Art. 13; high-risk via Art. 6 + Annex I; automation bias Art. 14(4)(b).

The explanation as the feedback signal — grounded, not speculative



Acting on output already pays

Predictive maintenance cuts maintenance cost 18–25 % and unplanned downtime up to ~50 % (McKinsey).

Explanations already correct models

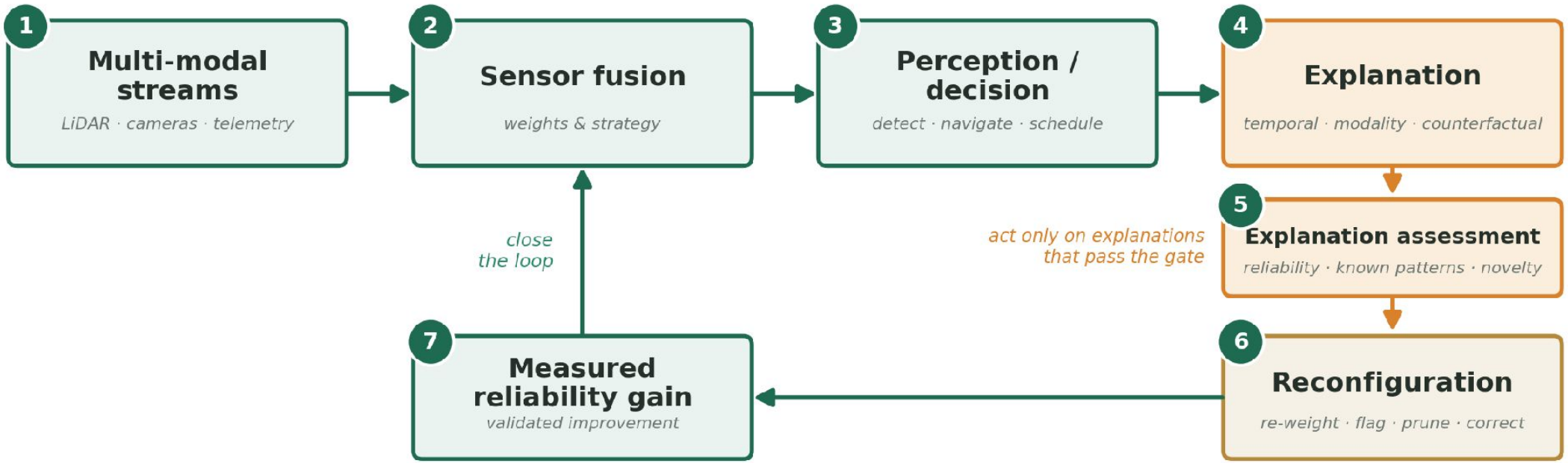
Explanatory interactive learning feeds explanation-level feedback into training (Ross et al. 2017; Schramowski et al. 2020).

The methods are mature

Attribution, temporal & counterfactual XAI are established (Guidotti et al. 2018; Barredo Arrieta et al. 2020).

PdM: McKinsey, “Prediction at scale”. Loop lineage: Ross et al. 2017; Schramowski et al. 2020; MAPE-K: Kephart & Chess 2003.

Walking the loop



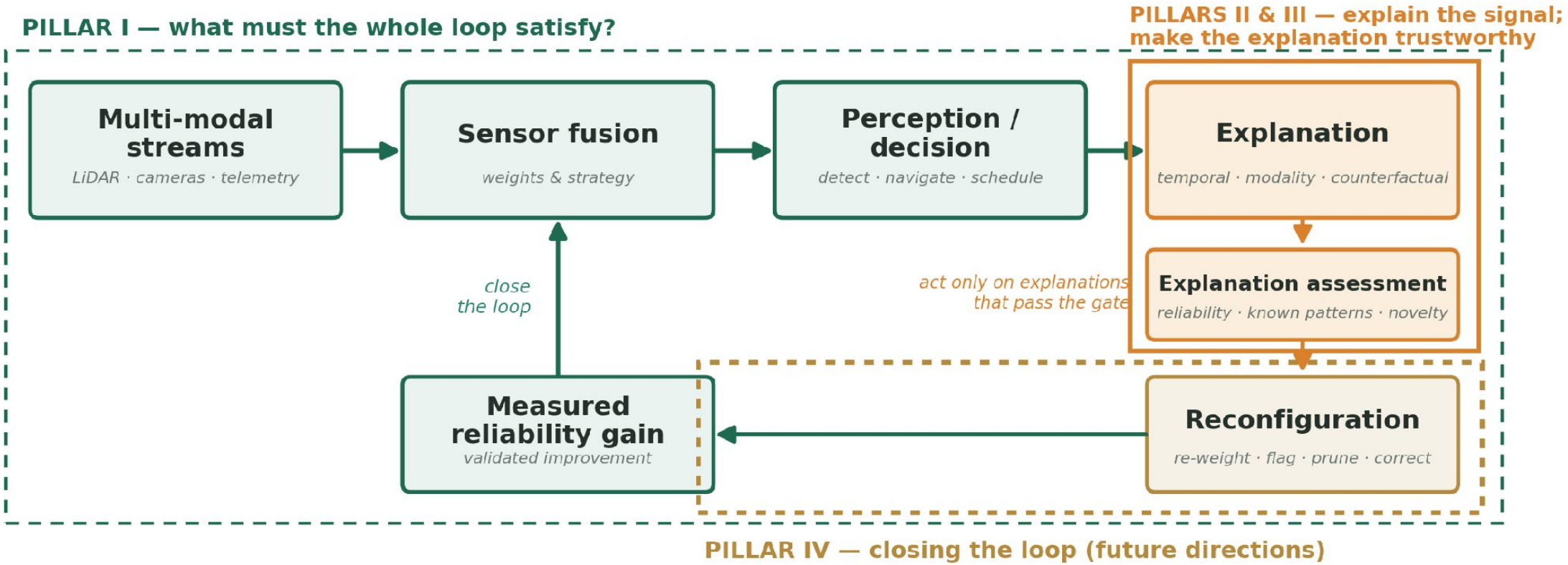
the explanation is the feedback signal — not the output

A generic anatomy — no research of ours in this picture yet.

Loop scheme: Polish–Taiwanese International Scientific Seminar, June 2026 (M. Mishchuk) — extended here with an explicit explanation-assessment step.

• AGENDA

Today: four pillars on one loop



The question, and how we examined it

What must be true for explanation → action to be valid — trustworthy enough to act on, and certain that acting helps — and is there a complete, closed set of such conditions?

Reasoning by precedent
case-based / analogical

Mature fields where a read-only signal became a control signal — governor, control chart, artificial pancreas (backup).

→ shows the move is real

Conditions-of-possibility
first-principles

What must hold for the move to work at all, independent of history.

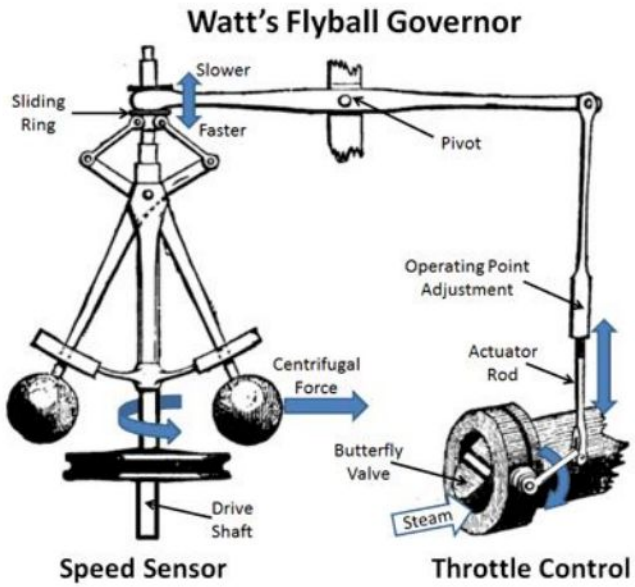
→ gives completeness

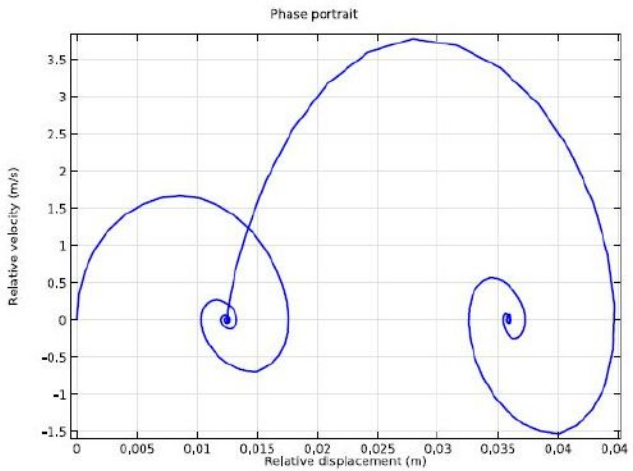
Dialectical analysis
thesis–antithesis–synthesis

Each condition is genuinely contested — so we earn a position, not a slogan.

→ a defensible synthesis

Once you feed a signal back into the thing it measures, what new problem appears?





Phase portrait: trajectories spiral in to equilibrium — a stable loop.

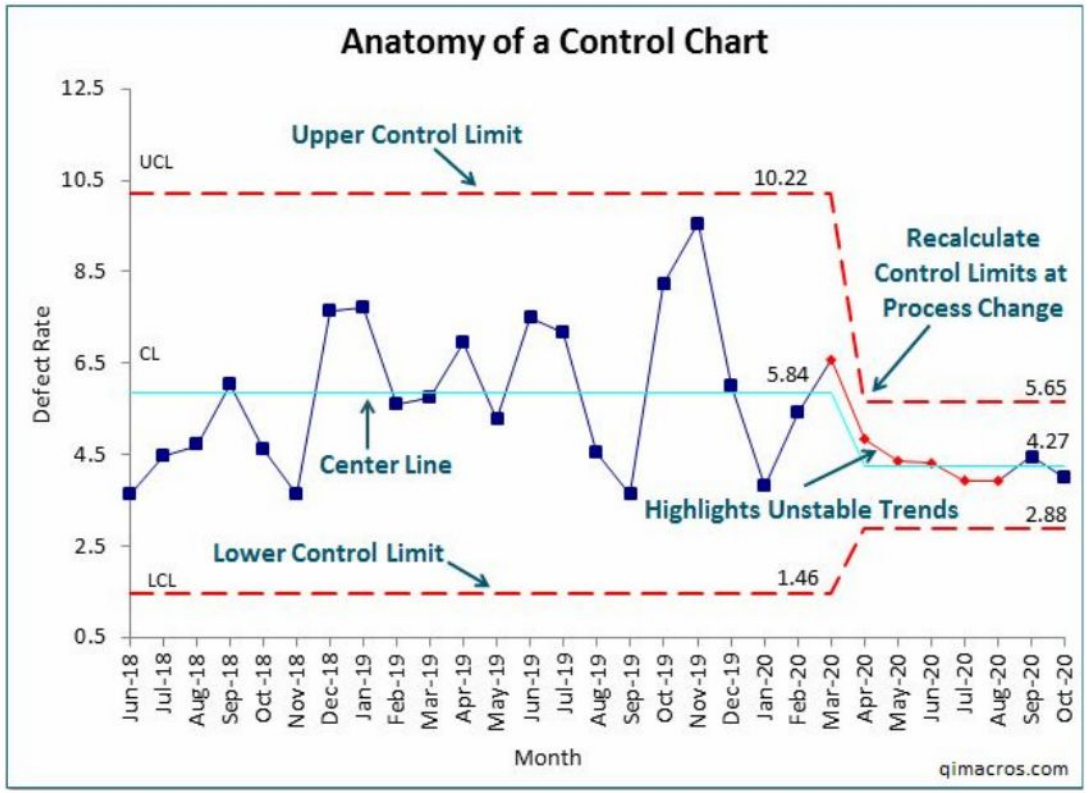
- Flyweights rise as the engine speeds up, throttling the steam valve as they rise.
- Engine speed is fed back to regulate engine speed — the first famous closed loop (Watt, ~1788).
- But governors could "hunt" — oscillate instead of settling.

THE LESSON

Maxwell, On Governors (1868), made feedback into mathematics: a closed loop must be stable — disturbances must decay, not grow.

Watt centrifugal/flyball governor (~1788). Maxwell, On Governors, Proc. R. Soc. Lond. 16:270–283 (1868).

Before a signal may trigger action, what must you be able to tell about it?



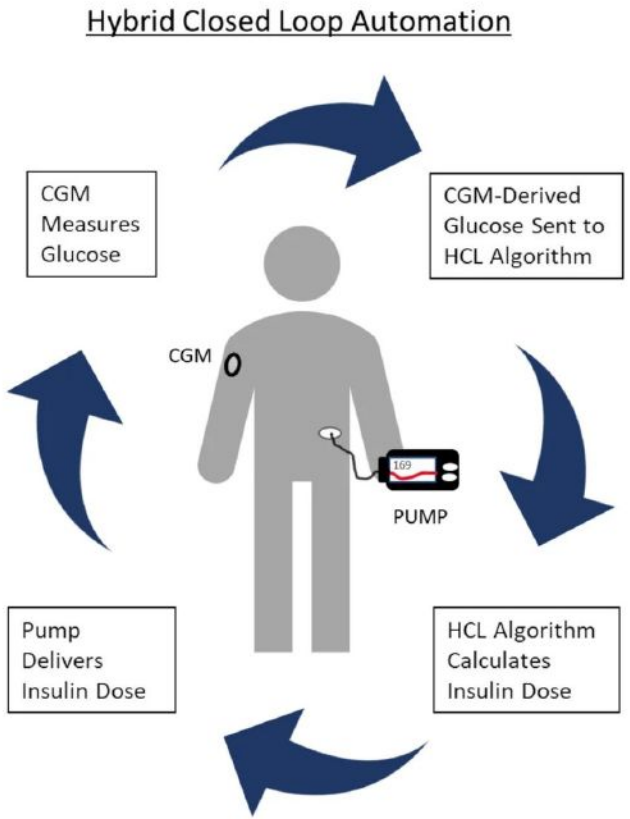
- Bell Telephone Labs / Western Electric, 1924 — making telephone gear that, often buried underground, had to fail rarely.
- Shewhart's chart separated common-cause (chance) noise from special-cause (assignable) signal, using ±3σ control limits.
- It told engineers when to intervene — and when to leave the process alone.

THE LESSON

You cannot act on a signal you can't first trust. Separating signal from noise is the precondition — it became closed-loop quality control (1931).

Shewhart, Bell Telephone Labs / Western Electric (memo, 16 May 1924); Economic Control of Quality of Manufactured Product, Van Nostrand (1931).

When a number a human used to read acts on its own, who holds the controls?



- Continuous glucose monitoring (CGM) was, for years, a readout clinicians and patients interpreted.
- In 2016 the first hybrid closed-loop system let the CGM signal modulate insulin delivery directly (FDA).
- Note the word “hybrid” — the controller handles fast routine corrections; the person retains the high-stakes ones.

THE LESSON

Closing the loop didn’t remove the human — it forced an explicit allocation of authority: who acts, and when.

Medtronic MiniMed 670G — first hybrid closed-loop system; US FDA approval, 28 Sep 2016.

First principles: the stages of turning a signal into control

STAGE 1
acquire & trust

not faithful → act on a falsehood

FAITHFULNESS

STAGE 2
decode into meaning

not localised → nothing definite to act on

LOCUS

STAGE 3
use by an agent

no warranted agent → action mis-assigned

AGENCY

STAGE 4
embody a sustaining loop

not stable / verifiable → can't run or prove it helps

REALITY

Three independent traditions — control theory, robotics' sense–plan–act, and autonomic computing's MAPE-K — converge on the same staged anatomy. There is no fifth stage.

Staged feedback loop: Maxwell 1868; Åström & Murray 2008; sense–plan–act: Fikes & Nilsson 1971 (STRIPS); MAPE-K: Kephart & Chess 2003.

Four necessary conditions

1 Faithfulness
Can we trust the explanation enough to act on it?

2 Locus
What — and at which level — do we explain?

3 Agency
Who closes the loop — human or machine?

4 Reality
Can the loop run, stay stable, and be shown to help?

Remove any one, and the explanation → action arrow breaks.

The three precedents instantiate exactly these four — and the first-principles decomposition derives the same set.

Complete relative to the stages — corroborated, not a theorem

What it is

An architectural / structural completeness argument — not a formal theorem.

A fifth condition would have to name a fifth stage in "signal → control."
I invite that challenge.

Two converging supports

(i) the staged decomposition — one failure mode per stage;

(ii) three mature loops that each solved exactly these and nothing more.

Two cross-cutting threads — stability (inside Reality) and a "memory of what correct looks like" (across all four) — are named as threads, not separate conditions.

Each condition is a live battlefield

Both poles are valid — solutions must navigate between them.

1 FAITHFULNESS	post-hoc attribution (LIME · SHAP · attention)	↔	inherent interpretability; disagreement problem	**OPEN GAP** validation without ground truth
2 LOCUS	one global feature ranking	↔	multi-level: temporal · modality · decision	**OPEN GAP** unified instance-level fusion explanation
3 AGENCY	human oversight (EU AI Act Art. 14)	↔	machine self-adaptation — too fast to supervise	**OPEN GAP** certifying a partly self-modifying loop
4 REALITY	explanation as a free post-hoc layer	↔	edge · energy · stability limits	**OPEN GAP** benchmark for explanation-driven control

Dialectical analysis per condition: Polish–Taiwanese seminar, June 2026 (full thesis–antithesis–synthesis slides in backup). Cross-cutting thread: a memory of what "correct" looks like.

Past work, in two lines

A INFRASTRUCTURE
Simulation, benchmarking, data pipelines, edge deployment groundwork → documented in TUAI Deliverable 3.1.

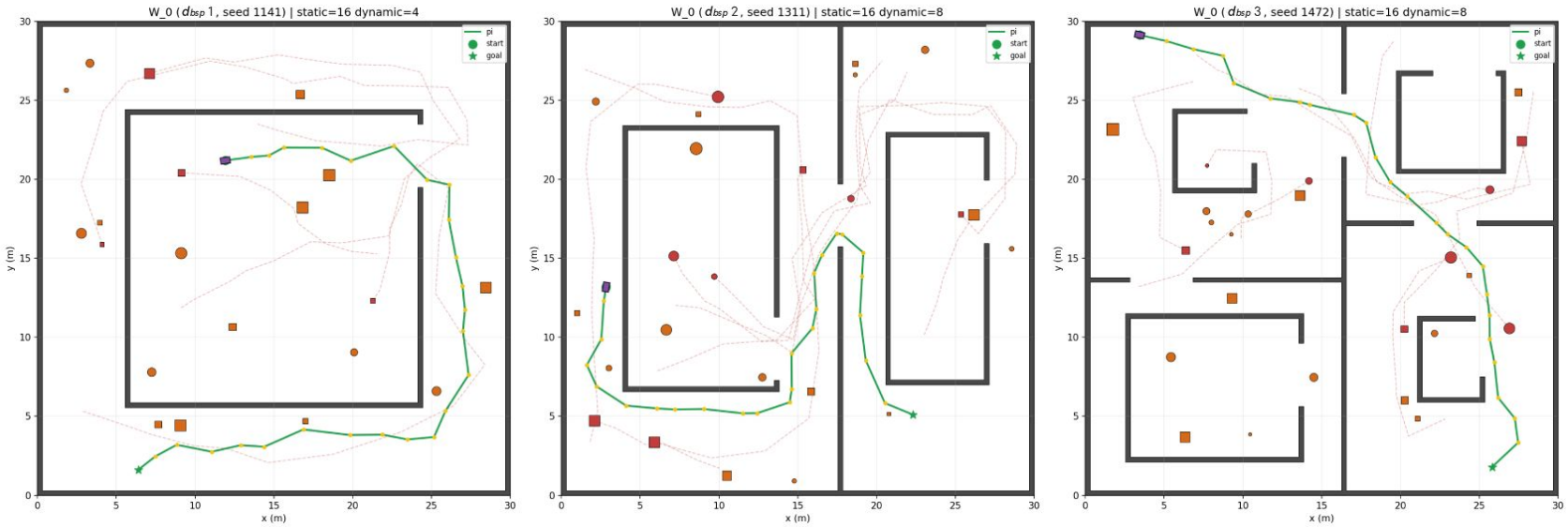
B PUBLICATIONS — EXPLAINING INDIVIDUAL SIGNALS
One sensor at a time, with simple explanations — the basis for everything that follows. → next slides.

Year one

#	Title	Venue	Authors	Status
1	KG-SEA: Self-Evolving Framework for Iterative Knowledge Graph Construction in Graph-RAG Systems	IEEE BigData 2025	Chen, Mishchuk, Niyomugaba, Lin, Cupek	**Published**
2	Iterative Weighted-Voting Approach for Adaptive Time-Series Anomaly Correction in Federated AGV Systems	IEEE BigData 2025	Pavliuk, Mishchuk	**Published**
3	Explaining LSTM Battery RUL Prediction via Temporal Attribution	ACIIDS 2026 · LNCS 16530	Mishchuk, Cupek, Pavliuk	**Published**
4	Energy-Aware Spatio-Temporal Multi-Agent Route Planning for AGVs	Applied Sciences 16(6)	Pavliuk, Mishchuk	**Published**
5	Simulation-Driven Benchmarking Framework for LiDAR-Based Collision Avoidance on AGVs	FLICS 2026	Mishchuk, Schiavo, Barani, Huda, Cupek	**Accepted**
6	LiDAR-Augmented Wi-Fi Signal Strength Prediction: A Use-Case Study on Industrial AGVs	FLICS 2026	Mishchuk, Pavliuk, Smołka, Tesliuk, Cupek	**Accepted**
7	HySDG-ESD: Hybrid Static-Dynamic Geometric LiDAR-Based Extended State Descriptor for AGV Obstacle Perception	IEEE SWC 2026	Barani, Huda, Mishchuk, Schiavo	**Accepted**
8	Personalized Worker Physiological Load Assessment Using Multimodal Wearable PPG Analysis and Activity Recognition	IEEE Access	Litovska, Mishchuk, Pavliuk	Submitted (preprint)

First-year report (Mishchuk, 2026), publications table. DOIs in the report and the papers folder.

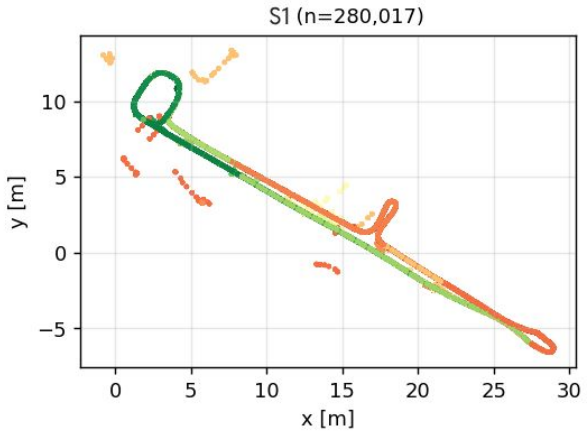
LiDAR

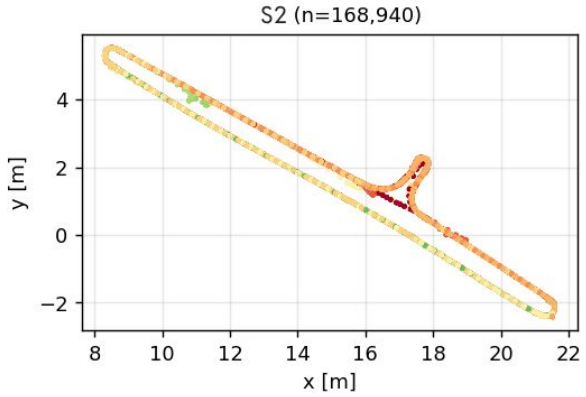


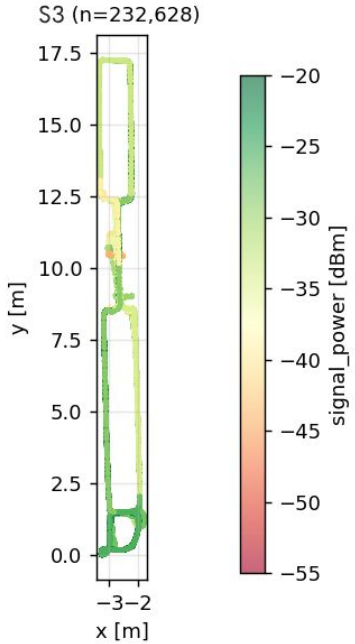
FLICS 2026

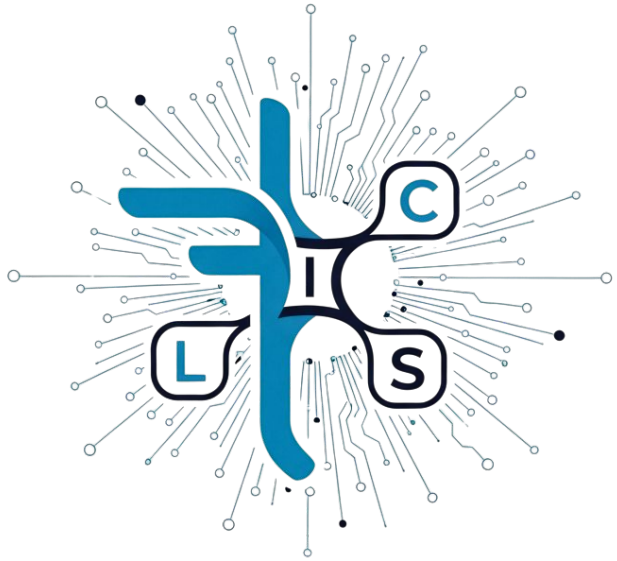
Simulation-driven benchmarking of LiDAR-based collision avoidance on AGVs.

Wi-Fi





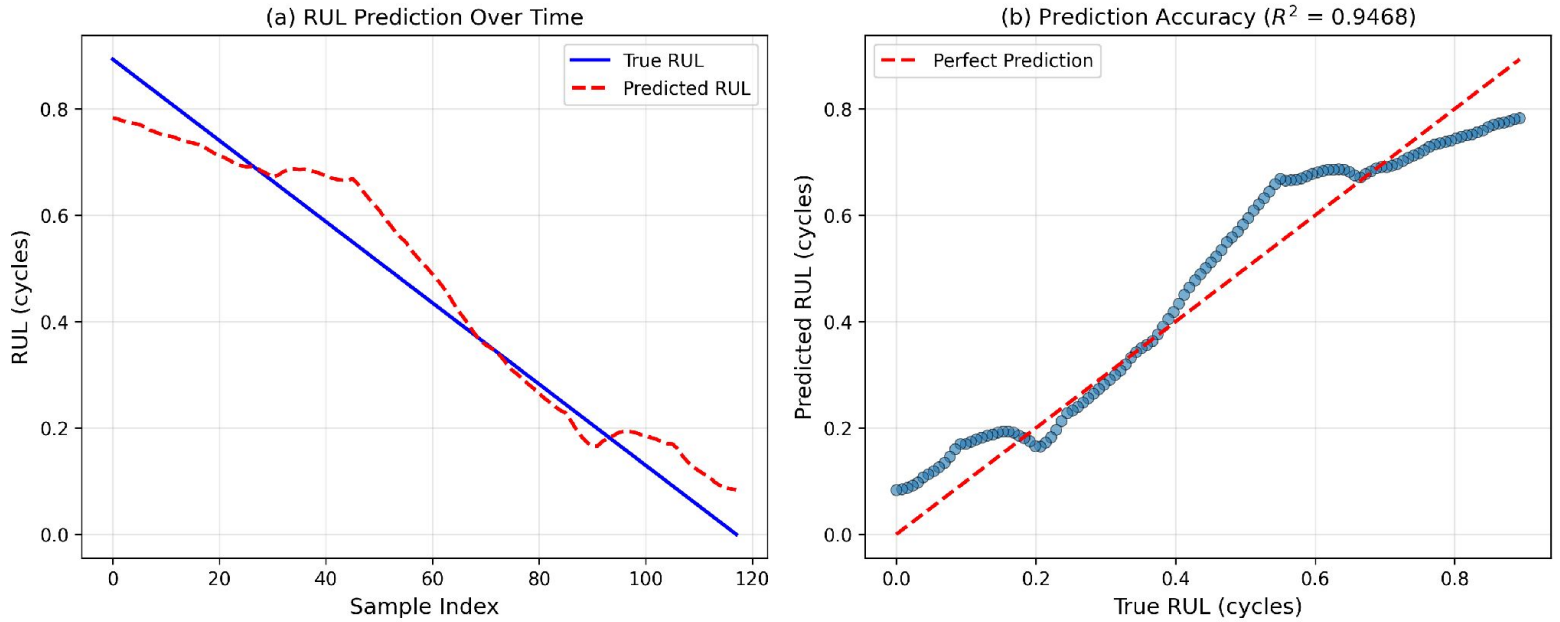




FLICS 2026

Wi-Fi signal-strength prediction — explained with SHAP and ablation.

Battery

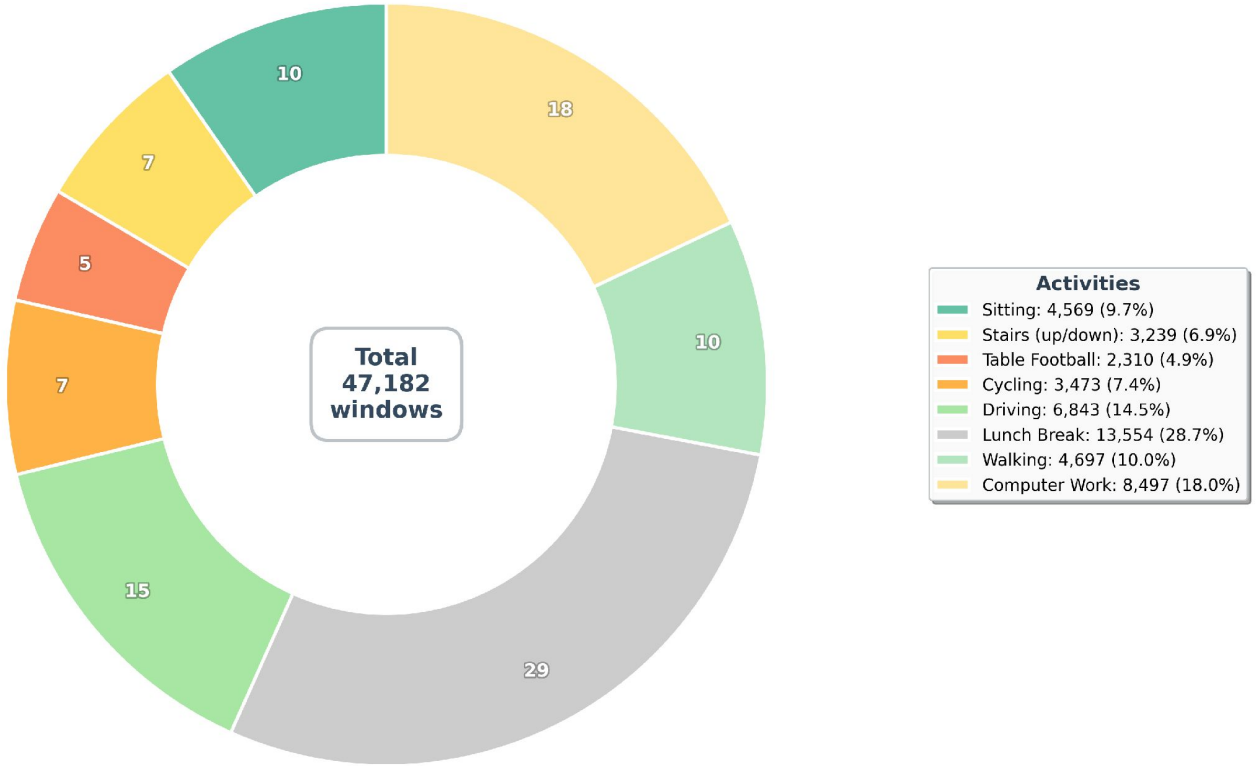


ACIIDS 2026

ACIIDS 2026

Battery RUL prognostics with an attention-LSTM — explained via temporal attribution (TimeSHAP).

Motion & wearables

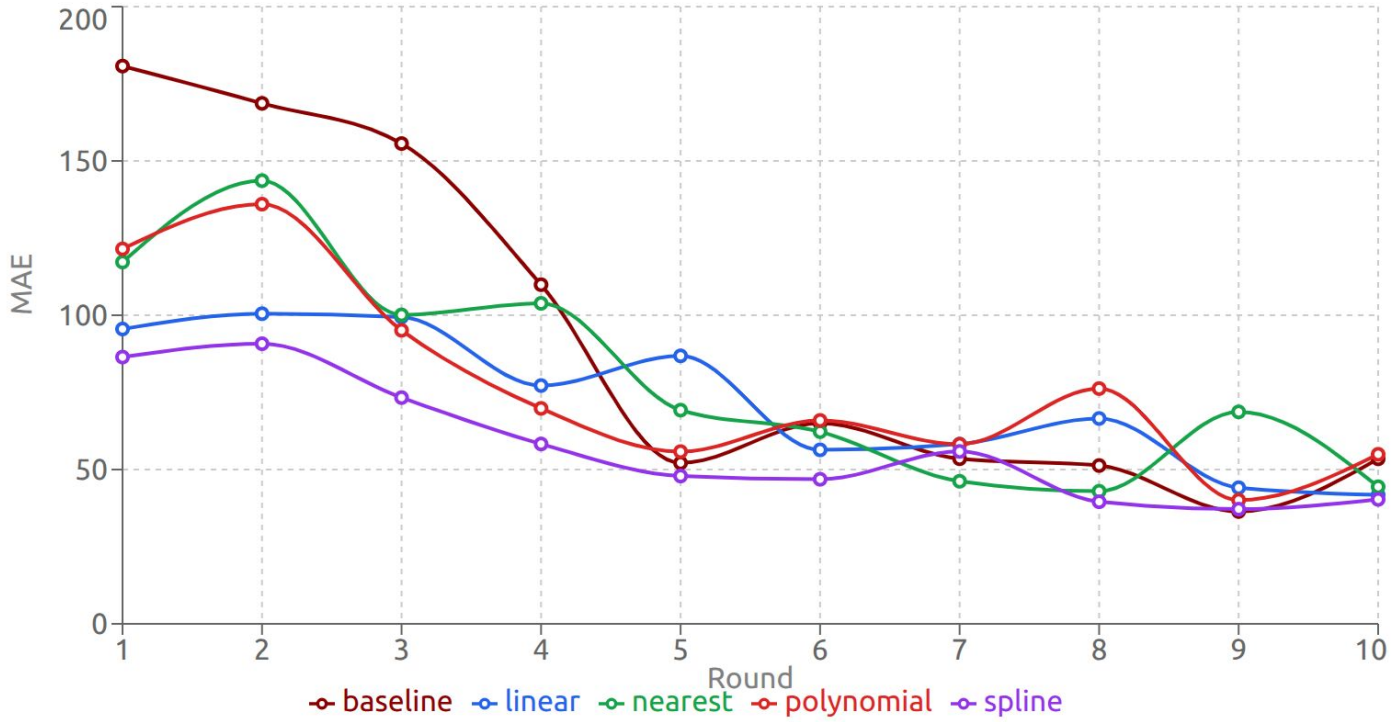


IEEE Access®

Preprint

Worker physiological-load assessment — wearable PPG + activity recognition, SHAP-based feature importance.

…and the fleet level

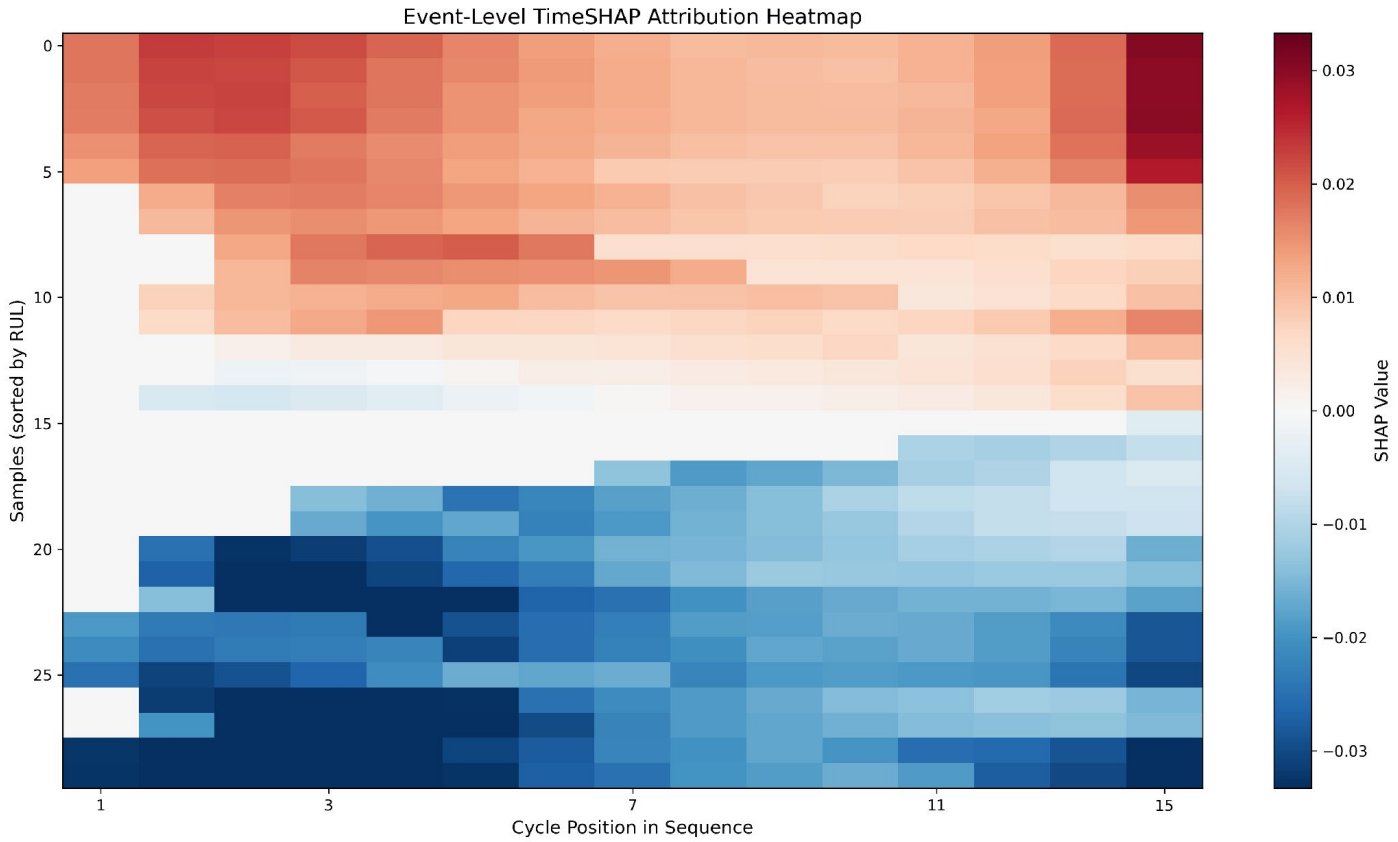


IEEE BIG DATA 2025

BigData 2025 · MDPI Appl. Sci.

Federated anomaly correction across rounds (~25 % error reduction) · energy-aware multi-agent routing.

But is this explanation trustworthy?



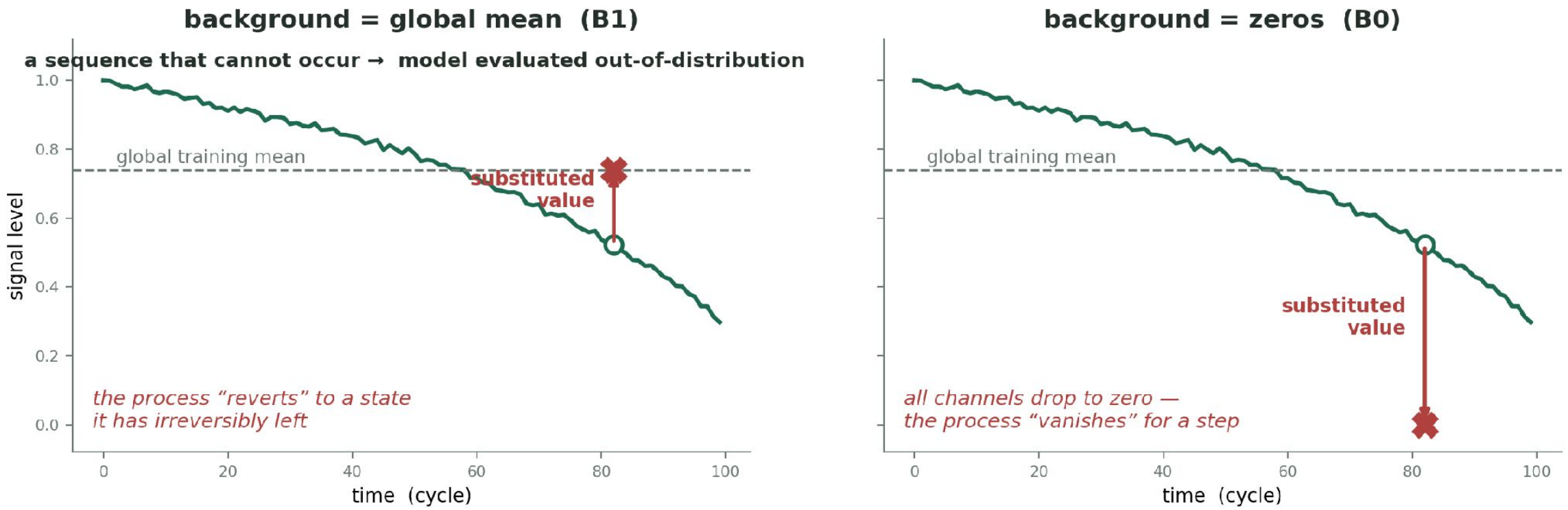
What we published:
attributions sharp at the edges — recent cycles strongest, oldest cycles moderate — mild in the centre. “The LSTM memorizes the first block as a baseline reference.”

The question:
is that the model — or an artifact of how the attribution was computed?

That gap is the first part of my current research — the field is moving the same way: from producing explanations to ensuring they are reliable.

Event-level TimeSHAP attribution heatmap: Mishchuk et al., ACIIDS 2026, Fig. 2.

The mechanism: masking means substitution



The attribution sums genuine information loss + the model's arbitrary reaction to an impossible input — in one non-decomposable number.

TimeSHAP background interface: Bento et al., KDD 2021. Defaults (zeros / global mean): field practice; our ACIIDS pipeline used the global feature-wise mean (Eq. 3).

“The explainer assumes stationarity where the model does not.”

The trend level is signal — the model legitimately consumes it.
An unconditional background silently imports a stationarity assumption the data violates.

(And no — detrending is not an escape: post-hoc XAI must explain the deployed model as it is; detrending at explanation time recreates the same out-of-distribution mechanism.)

Organizing claim of the methodology (Masking Backgrounds v5, §1.1); detrending argument: §1.4.

Six hypothesized artifacts, three classes

CLASS I — MAGNITUDE

attribution size driven by geometry, not information

A1 Edge inflation
extremes lie far from the mean → inflated attributions

A2 Mid-trajectory blindness
near-mean steps barely perturbed → squashed to zero

CLASS II — SEMANTIC

the substitute means something to the model

A3 Spurious-excursion response
masking manufactures a fake "regeneration" event

A4 Directional offset
late windows pushed toward earlier trend positions

CLASS III — PROPAGATION

single-step distortion spreads

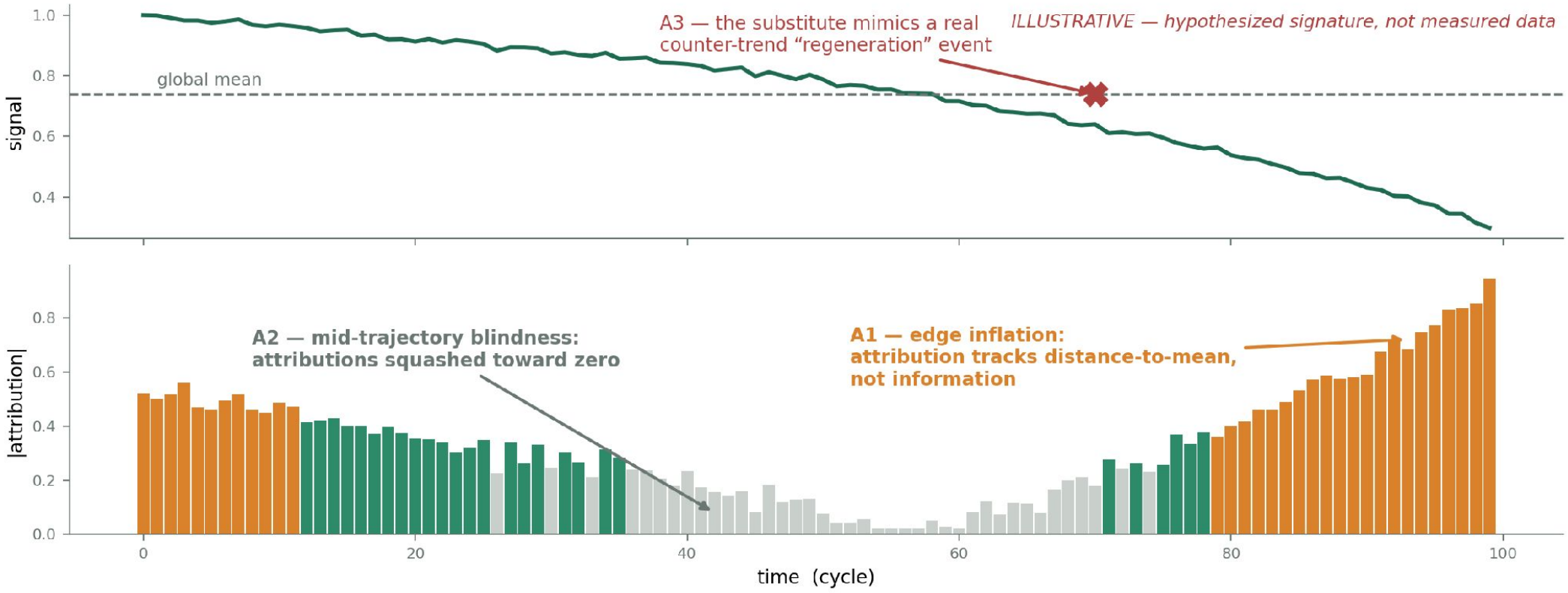
A5 Recurrent state carry-over
corrupted hidden state pollutes all later steps

A6 Coalition compounding
multi-step masking compounds A1–A5 per call

Sharp-knee regimes act as an amplifier of A1 / A3 / A4 — not a seventh artifact.

A-taxonomy: Masking Backgrounds methodology v5, §1.2.

The artifact signature — and a familiar pattern



Edge-sharp, centre-mild — exactly the ACIIDS pattern. Our published result is currently indistinguishable from these artifacts.

Illustrative panel (hypothesized signature, not measured data). ACIIDS limitations & Reviewer 4 anticipated the weak masking baseline; Masking Backgrounds v5, §1.3.

Four literatures near the problem — none at the intersection

Tabular Shapley — off-manifold
conditional / generative value functions; static data, generic manifolds

Image XAI — masking artifacts
missingness bias; fixes need an inpainting model; no temporal ordering

THIS WORK
harder problem, cheaper fix

Time-series saliency (non-Shapley)
realistic perturbations, but abandons Shapley axioms; (quasi-)stationary benchmarks

Temporal Shapley methods
background sensitivity measured — but all candidates remain unconditional

Missing everywhere: (i) characterization of trend-structured bias · (ii) conditional backgrounds for mean-nonstationary data · (iii) any ground-truth benchmark with sustained, label-relevant mean-nonstationarity.

Prior-art map: Masking Backgrounds v5, §2 (novelty searches re-run 2026-07-19 — no direct competitor on trend-/state-conditional backgrounds for temporal Shapley).

Three rules a TAB background must obey

`b(t, x)` returns one replacement value per channel — a vector of length C

R1

Past only

No leakage

Uses x[1..t] and label-free training statistics — nothing else.

Forbidden: normalized position t/T. Computing T means knowing the remaining life.

Allowed: elapsed age t.

R2

Computed once

Before any coalition

b is computed from the original observed sequence, never recomputed from a masked one.

Keeps the value function a single well-defined set function — and makes the cost negligible.

R3

No shape assumption

Nothing about the trend

No monotonicity, no linearity, no parametric trend form.

Isotonic or monotone-spline fitting enters only as a declared variant, reported separately.

Both TAB constructions are single-reference backgrounds, matching the one-background-instance interface of the estimator.

Trend-Aware Backgrounds: a ladder of conditioning

B0
zeros
incumbent default

B1
global mean
incumbent (ACIIDS Eq. 3)

B-L
carry-forward
local interpolation

OURS

B2
TAB-state
state-conditional population mean (past-only)

OURS

B3
TAB-traj
the sequence's own strictly-past trajectory fit

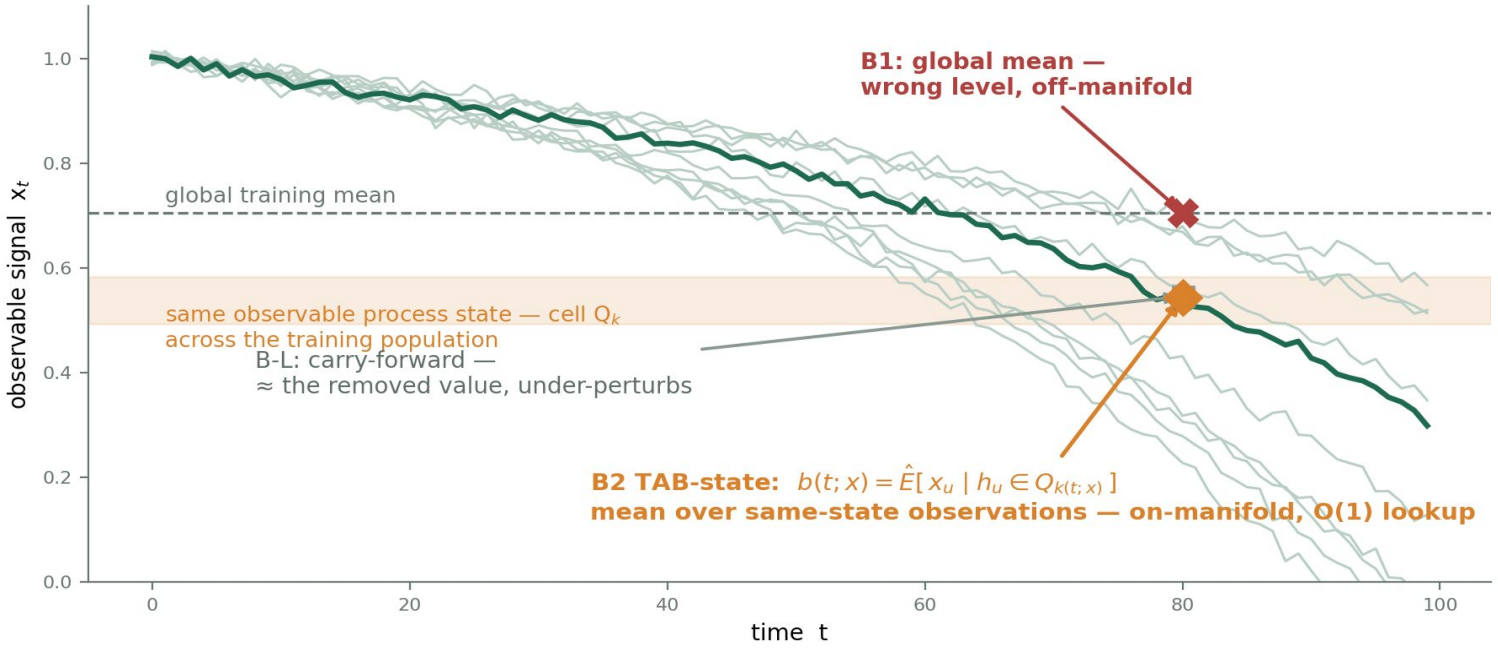
B4
learned imputer
generative reference arm — not an assumed ceiling

conditioning resolution → (B0/B1 are the degenerate members — the experiment varies conditioning and nothing else)

Near-free by design: precompute + lookup — no generative model to train. Leakage discipline: backgrounds use the past only (normalized position t/T is inadmissible; elapsed age is). B3 masks exactly the timestep's innovation.

TAB-state / TAB-traj formal definitions: Masking Backgrounds v5, §§3.3–3.4 (details in backup). Every ordering vs B4 is publishable.

TAB-state: the state-conditional expectation



SETUP

Sequence x ∈ ℝ^(T×C); a background is a map b: (t, x) ↦ b(t; x) ∈ ℝ^C, substituted at masked entries.

STATE INDEX & BINNING

h□(x) = h(x₁:□) ∈ ℝᵈ — past-only (F□-measurable); reference d = 2: elapsed age t + an observable state estimate ĥ□. Partition the range of h into K quantile cells Q₁ … Q_K on training data only.

DEFINITION & ESTIMAND

b_state(t; x) = mean of training observations whose state falls in the same cell — the binned estimate of E[x_u | h_u ∈ Q_k(t;x)].

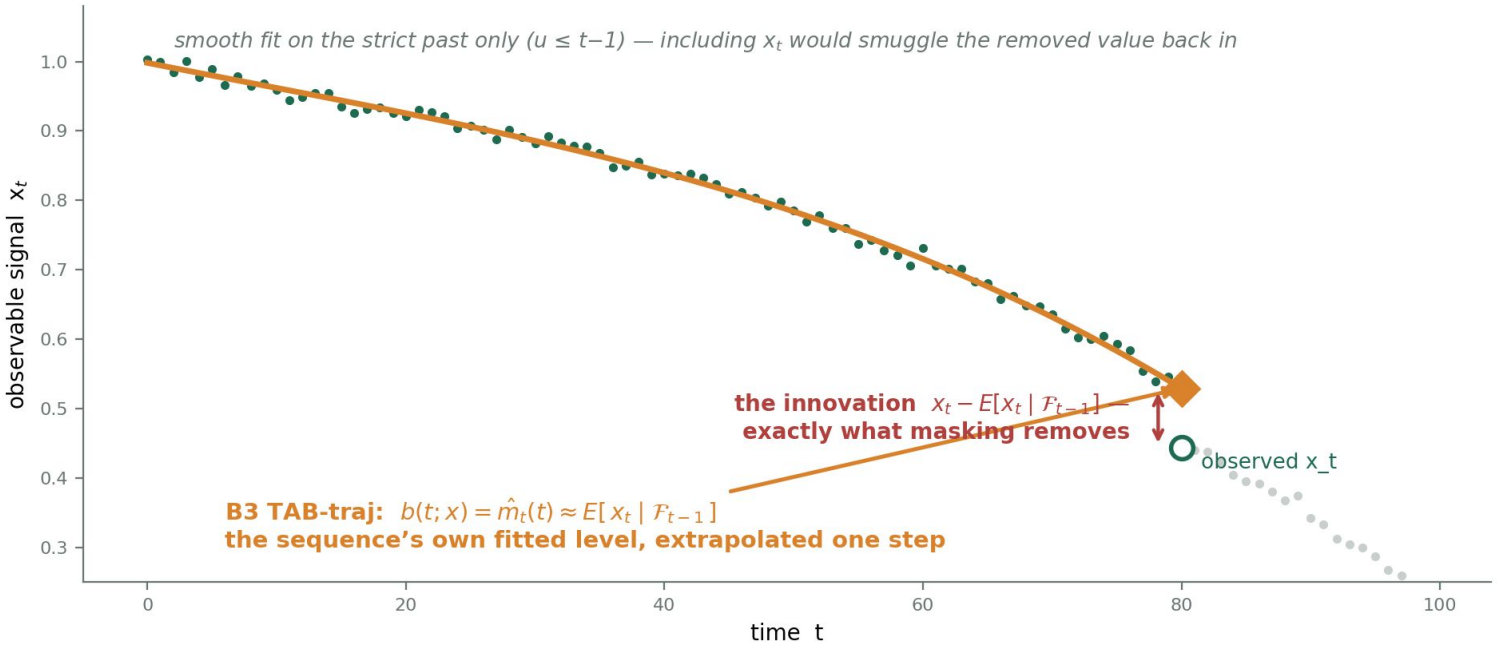
PROPERTIES

R1 past-only — t/T inadmissible (T encodes the label); R2 coalition-independent — one K×C table, O(1) lookup; R3 no shape assumption — valid under non-monotone μ(t). K = 1 recovers B1 exactly: a nested family.

vs B-L carry-forward: the substitute ≈ the removed value — under-perturbation, and importance goes blind.

Masking Backgrounds v5, §3.4.2. Panel: illustrative sketch of a mean-nonstationary process.

TAB-traj: the past-conditional expectation



FIT ON THE STRICT PAST

$\hat{m}_t$ = argmin over m ∈ M_fit of $\Sigma_{u \le t-1} w(t-u)\cdot(x_u - m(u))^2 + \lambda \cdot P(m)$
— a weighted, penalized smooth fit; family able to express sharp local curvature.

DEFINITION & ESTIMAND

b_traj(t; x) = $\hat{m}_t(t)$. Under $x_u = m(u) + \eta_u$ with $E[\eta_u \mid F_{u-1}] = 0$, this estimates $E[x_t \mid F_{t-1}]$ — so masking removes exactly the innovation $x_t - E[x_t \mid F_{t-1}]$.

WHY u ≤ t−1, NOT u ≤ t

Including x_t makes the substitute a smoothed copy of the removed value — B-L-style under-perturbation; it also breaks the estimand.

FALLBACK & VARIANTS (R1–R3 hold)

t−1 < t_min ⟶ pre-declared fallback to TAB-state; monotone/isotonic projection only as a declared per-profile variant, never the method (R3). Per-sequence fit residuals reported as a diagnostic.

vs B4 learned imputer: the established generative remedy needs a second trained model per dataset — cost, competence barrier, its own failure modes. B4 runs as a reference arm; every ordering is publishable.

Masking Backgrounds v5, §3.4.3. Panel: illustrative sketch of a mean-nonstationary process.

A benchmark that can measure this at all

RQ1 characterize | **RQ2 fix** | **RQ3 transfer**

CALIBRATED SYNTHETIC PROFILES

established generator pattern of the field

- double-exponential fade family (He et al. 2011)
- documented “knee” regime (Attia et al. 2022)
- planted ground truth — relevance known by construction
- **trap channels:** full trend geometry, zero label relevance — any attribution mass on them is distortion by definition

calibration →

← transfer (RQ3)

PAIRED REAL ARMS — BATTERY RUL

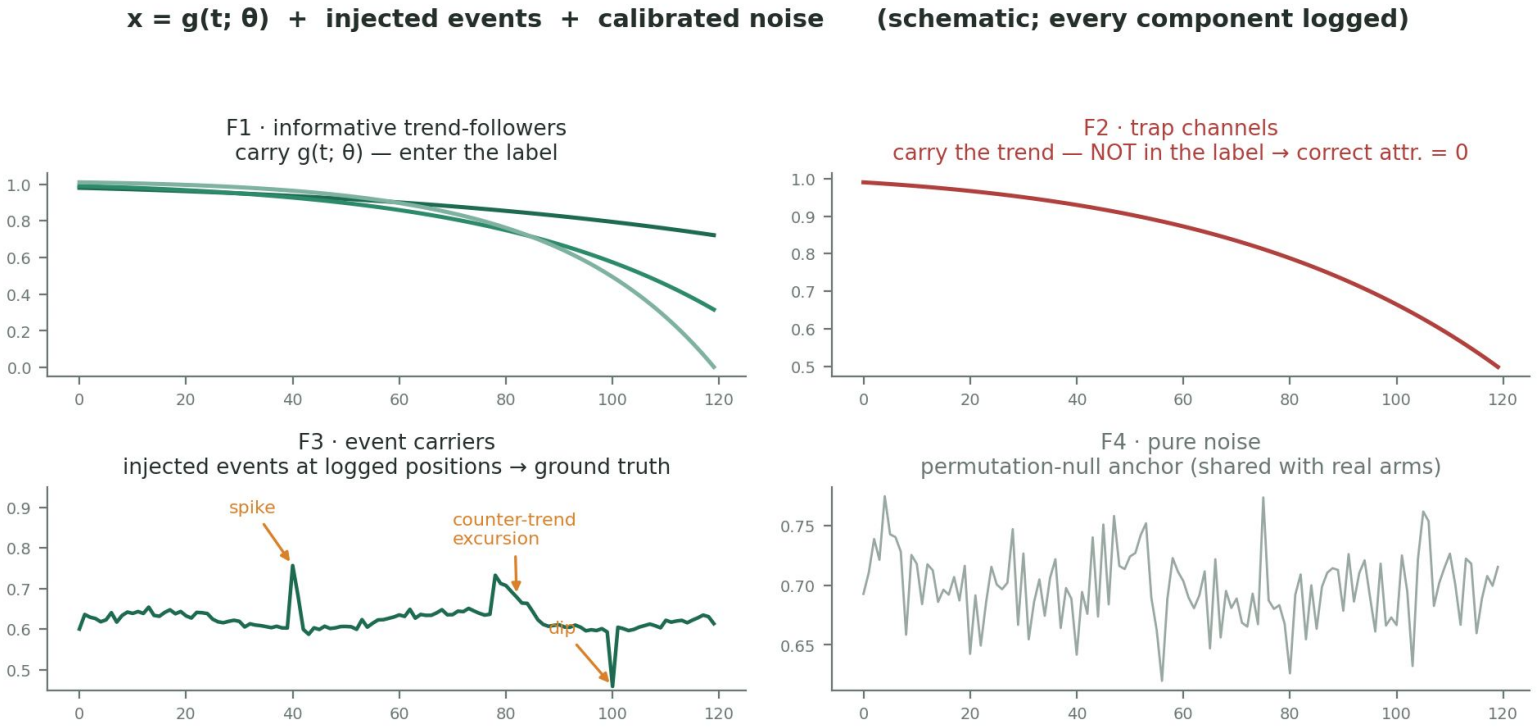
every stress condition exists in nature

- NASA PCoE — knee + documented regeneration excursions
- CALCE — second lab protocol
- Severson / MATR — statistical power, flat early trajectories
- NASA RW — rough, load-driven transients
- (optional) C-MAPSS — different physical domain

No existing time-series attribution benchmark provides ground truth under sustained, label-relevant mean-nonstationarity — the regime where the bias is structured has never been measurable.

Benchmark design & pairing table: Masking Backgrounds v5, §§3.6–3.7 (backup). Generator pattern: Tonekaboni et al., NeurIPS 2020 lineage.

Inside the synthetic arm



PLANTED GROUND TRUTH

The label is an explicit, logged function of specific cells — Design 1: anchors · Design 2: events · Design 3: the trap (label ignores F2 entirely).

CALIBRATED, NOT INVENTED

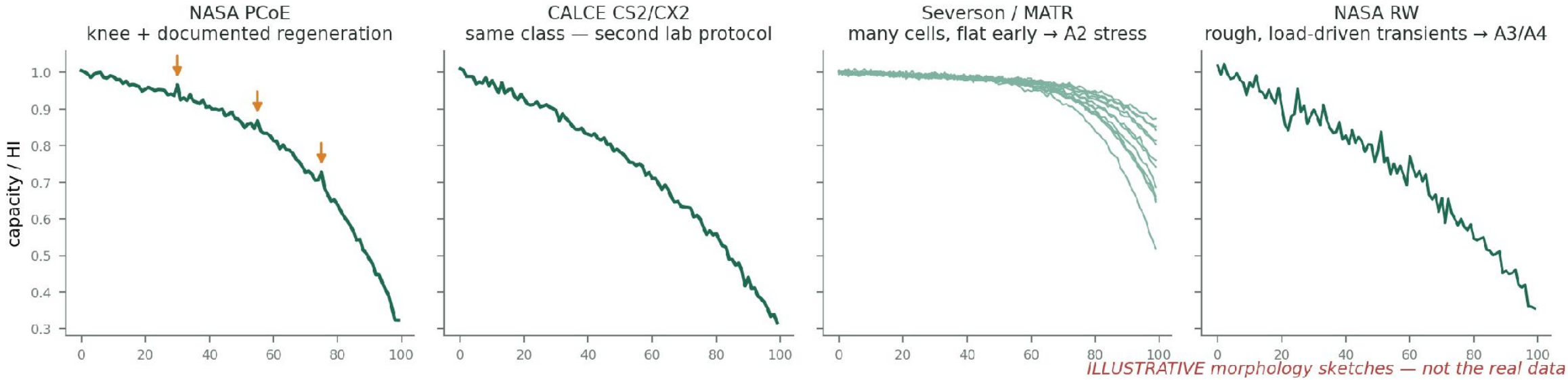
Trend family: double-exponential fade (He et al. 2011); θ-distribution, noise, sequence lengths and excursion statistics fitted to the paired real dataset; sharp-knee stress variant per profile (knee: Attia et al. 2022).

VALIDITY GATE

Attribution runs only after the model passes a near-ceiling accuracy gate — ground truth is meaningful only if the model learned the generating function.

Generator design & calibration recipe: Masking Backgrounds v5, §§3.6.1–3.6.3. Established pattern: Tonekaboni et al., NeurIPS 2020 lineage. Panels: schematic.

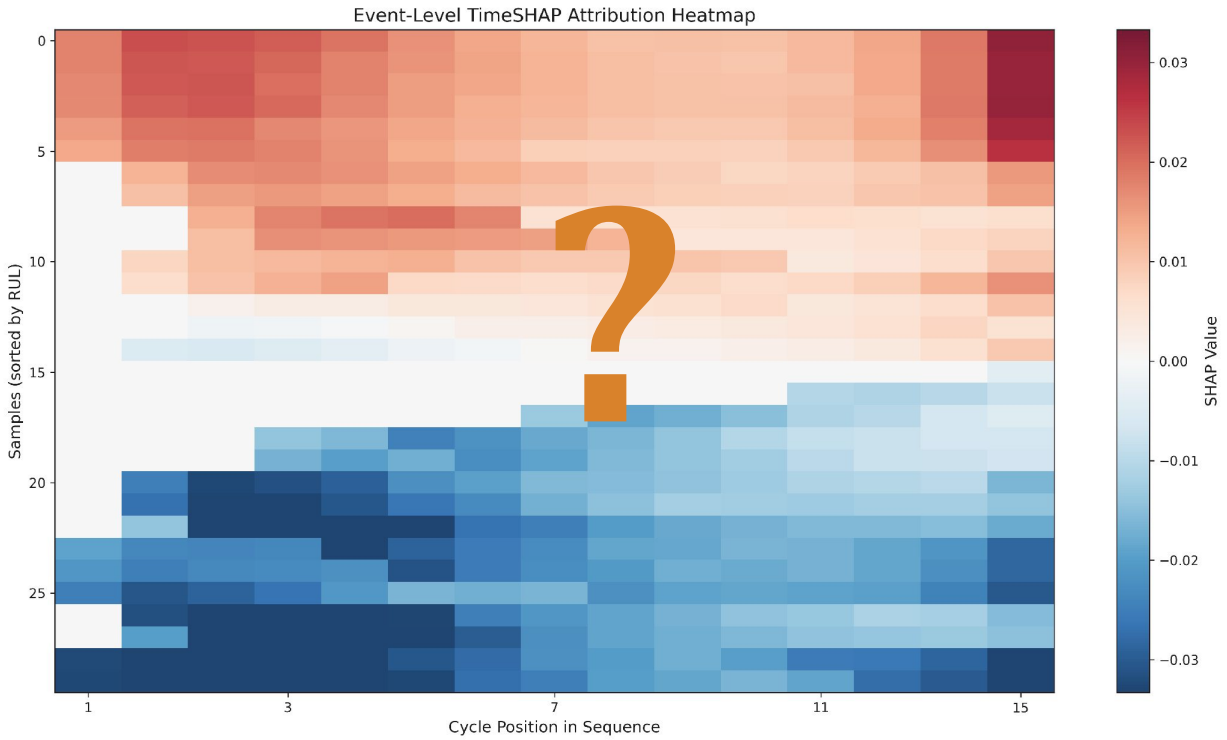
The paired real arms — one stressed property each



RQ3 — transfer: do the background rankings and artifact signatures found against synthetic ground truth reproduce on the real arms under proxy instruments? Plus P-GEN — a synthetic-only scope probe with non-monotone mean-nonstationarity — and, optionally, C-MAPSS for a different physical domain.

Pairing table & dataset property audit: Masking Backgrounds v5, §§3.6.4, 3.11. Sketches: illustrative morphology, not the real data.

Re-examining our own published result



The ACIIDS findings, re-produced on a corrected stack:

- corrected pipeline — leakage-free normalization, past-only features, retrained model
- corrected background — TAB instead of the global mean
- **the two fixes de-confounded — which conclusions survive?**

Status: methodology v5 finalized · novelty re-checked 2026-07-19 · experiments starting.

This is the point where your objections are cheapest — before the experiments run, not after.

Consequence case study: Masking Backgrounds v5, §§1.3, 3.14.

Closing the loop: an open design space

Once explanations are trustworthy — how does their feedback re-enter the system?
A vision, not a settled architecture.

1

Gated fusion re-weighting

bounded, certifiable, near-term

2

Explanation-guided adaptive replay

idea stage — explanations select what to learn next

…?

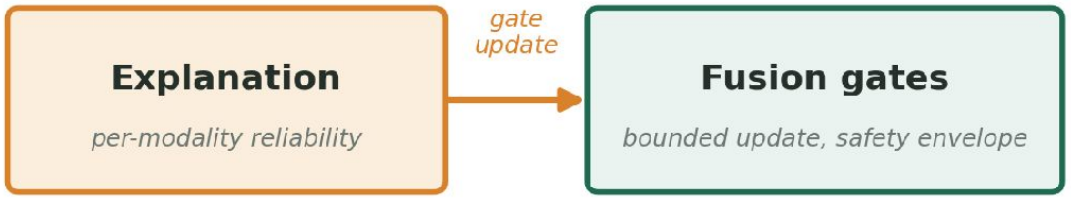
Your proposals

this fan is deliberately incomplete

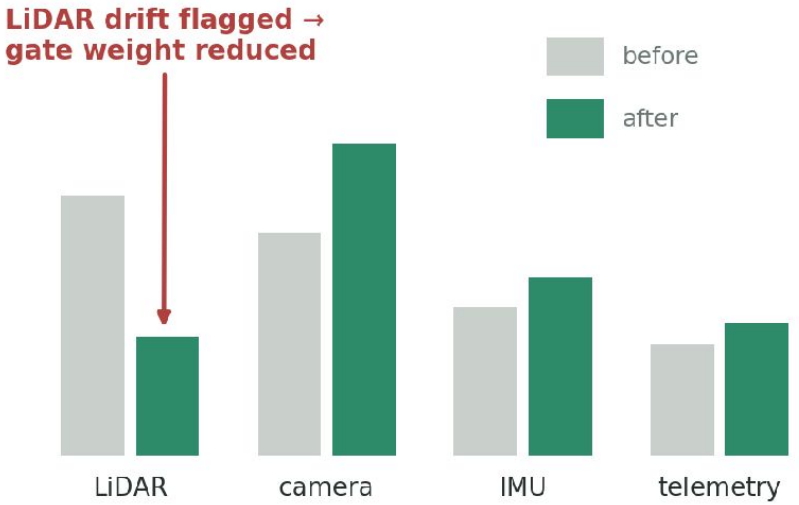
And whatever the mechanism: validating the applied change is a question of its own — after the pilot sketch.

Reconfiguration mechanisms — candidates, explicitly non-exhaustive.

Gated fusion re-weighting



human ratifies structural change · operators can always contest and override

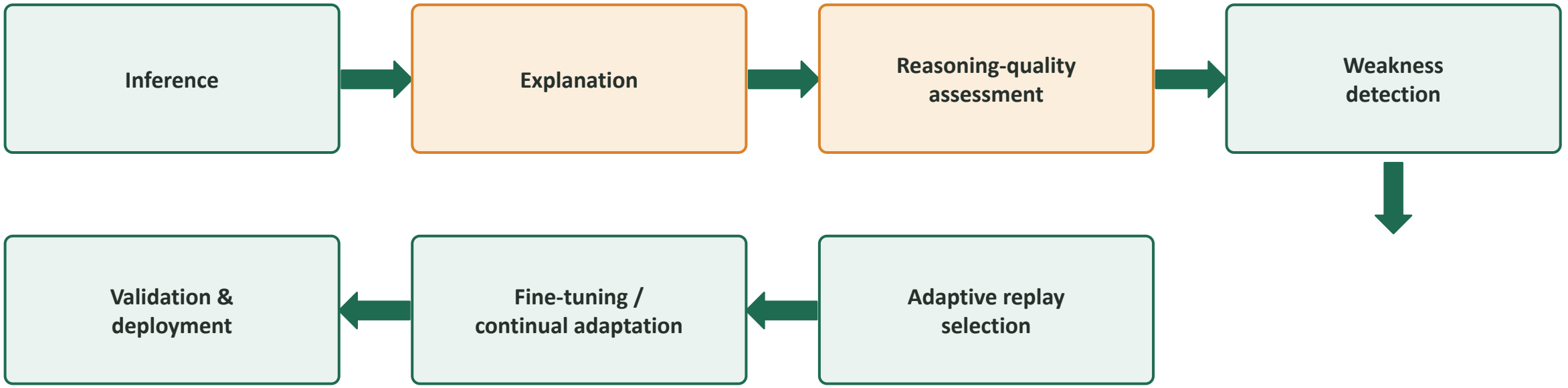


Existing evidence the loop family works: the federated anomaly-correction loop re-weights its own detectors (~25 % error reduction). Exposed requirement: a memory of what “correct” looks like.

Bounded-autonomy synthesis: Polish–Taiwanese seminar (Agency, p.25). Federated correction: BigData 2025.

Explanation-guided adaptive replay

Not "why did the model predict this?" — but "what do these explanations tell me about what I should learn next?"



No ground-truth labels assumed. Deployment signals — explanation stability, model confidence, sensor reliability, cross-sensor disagreement, temporal consistency — estimate reasoning reliability, not prediction correctness. The explanation selects what to learn; gradient descent decides how weights change.

Pilot design, and the long-term vision

ENVIRONMENT
NVIDIA Isaac Sim — AGV navigation & braking

GRADUAL DEGRADATION
LiDAR drift · camera noise · IMU drift · encoder slip · intermittent failures · lighting / weather

COMPARED STRATEGIES
1 static model
2 random replay
3 uncertainty-based replay
4 error-based replay (where labels exist)
5 explanation-guided replay

[Isaac Sim screenshot — suggested: NVIDIA developer media, credit on slide]

LONG-TERM
a step toward reflective AI: systems that identify their own reasoning weaknesses and choose what to revisit

Pilot & metrics: research note (backup). Avoids semantic weight editing; compatible with standard optimization.

After the change: does it help — and what risk did we accept?

THE QUESTION SET (unexplored)
We apply a reconfiguration — how do we know it works? Is the changed loop still stable, runnable, certifiable? This is the Reality condition, turned on our own change.

THE SIMPLEST ANSWER — AND ITS RISKS
Learn from operational data: compare operation stability before vs after the change. But: conditions drift between the two windows (confounded comparison) · a bad change does live damage while we measure · there is no counterfactual.

INITIAL THOUGHTS — FOR DISCUSSION

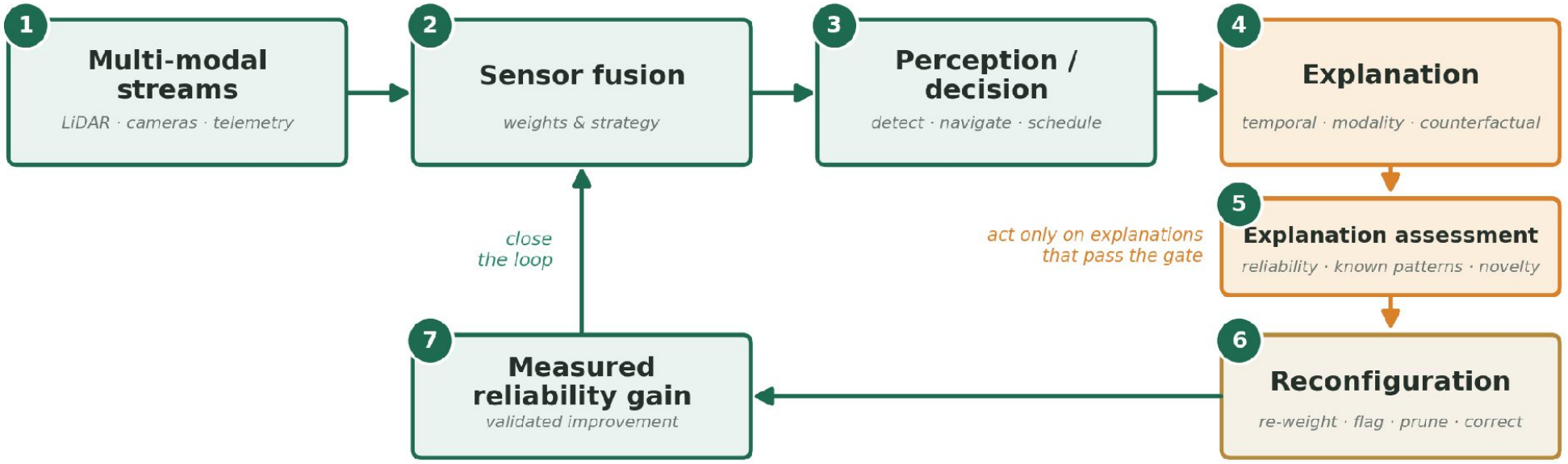
- Shadow / twin evaluation — let the change run in simulation or in parallel before it is allowed to act
- Staged rollout with control-theoretic safeguards — rate limits, hysteresis (the Reality synthesis)
- Control-chart the loop's own reliability metrics — Shewhart, back from pillar one: intervene only on special-cause change
- An explicit risk budget — how much deviation may a trial cost, and automatic rollback when it is spent

Status: future work — not yet investigated. Proposals and pointers to literature very welcome.

Reality-condition safeguards (rate limits, hysteresis): Polish–Taiwanese seminar, Reality synthesis. Control charts: Shewhart 1931 (precedent, backup).

• SYNTHESIS

One loop, three maturity levels



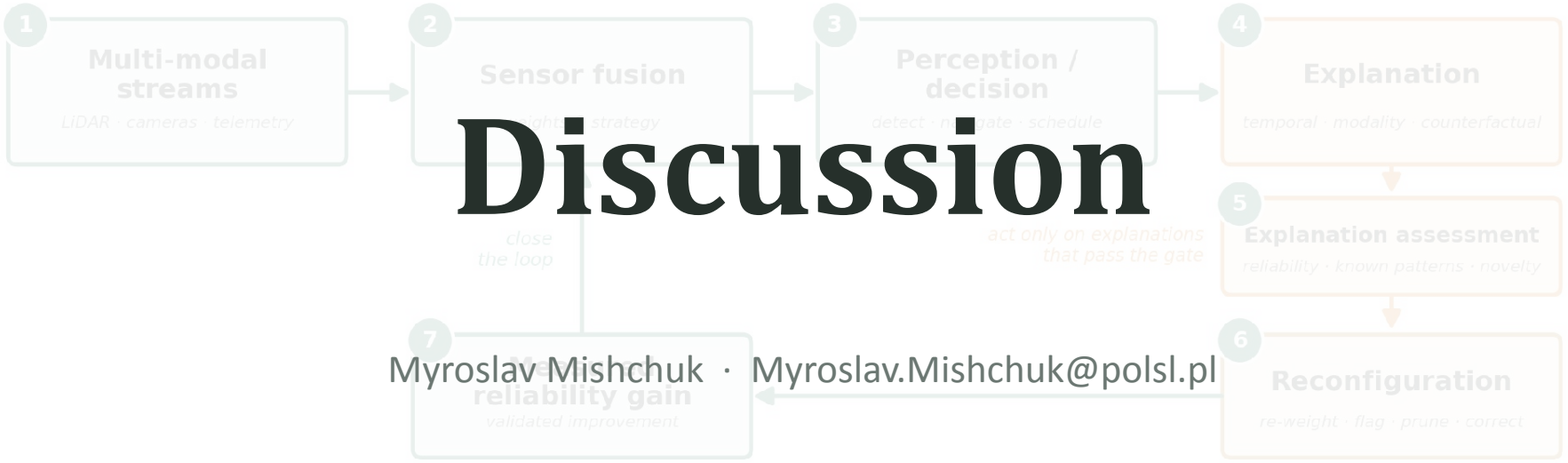
the explanation is the feedback signal — not the output

✓ signal explained — past work

• trustworthy explanations — now

… closing the loop — open

From explanation-as-interpretability to explanation-as-control — from the properties of that loop to its mechanisms.



Discussion

Myroslav Mishchuk · Myroslav.Mishchuk@polsl.pl

Backup

CONDITION 1 · FAITHFULNESS

Thesis: post-hoc methods make black boxes transparent

THESIS

LIME	A local surrogate linearises the model around a single prediction (Ribeiro et al. 2016).
SHAP	Game-theoretic Shapley values give an additive feature attribution (Lundberg & Lee 2017).
Attention	Attention weights read off as what a sequence model "attended to."

On this view, faithfulness is basically solved — pick a method and read it off.

Ribeiro et al. 2016 (LIME, KDD '16). Lundberg & Lee 2017 (SHAP, NIPS 2017).

Antithesis: “transparent” explanations can mislead

ANTITHESIS

- Build inherently interpretable models instead (Rudin 2019).
- Methods disagree with each other — the disagreement problem (Krishna et al. 2022).
- Post-hoc explanations can be adversarially fooled (Slack et al. 2020).
- Attention is contested: “not Explanation” (Jain & Wallace 2019) vs “not not Explanation” (Wiegreffe & Pinter 2019).
- Explanations breed over-trust even when wrong — automation bias (EU AI Act Art. 14(4)(b)).

REINFORCING CASE

A US NTSB report attributed a fatal 2018 automated-driving crash partly to over-reliance on automation, with monitoring that “facilitated the driver’s complacency.”

NTSB HWY18FH011 (2020)

Rudin 2019; Krishna et al. 2022; Slack et al. 2020; Jain & Wallace 2019 vs Wiegreffe & Pinter 2019; EU AI Act Art. 14(4)(b); NTSB HWY18FH011 (2020).

CONDITION 1 · FAITHFULNESS

Synthesis: gate action behind graded faithfulness

SYNTHESIS

(a)

Prefer inherently-interpretable / hybrid models where stakes are highest (Rudin).

(b)

Otherwise gate action behind graded faithfulness metrics + cross-method agreement.

(c)

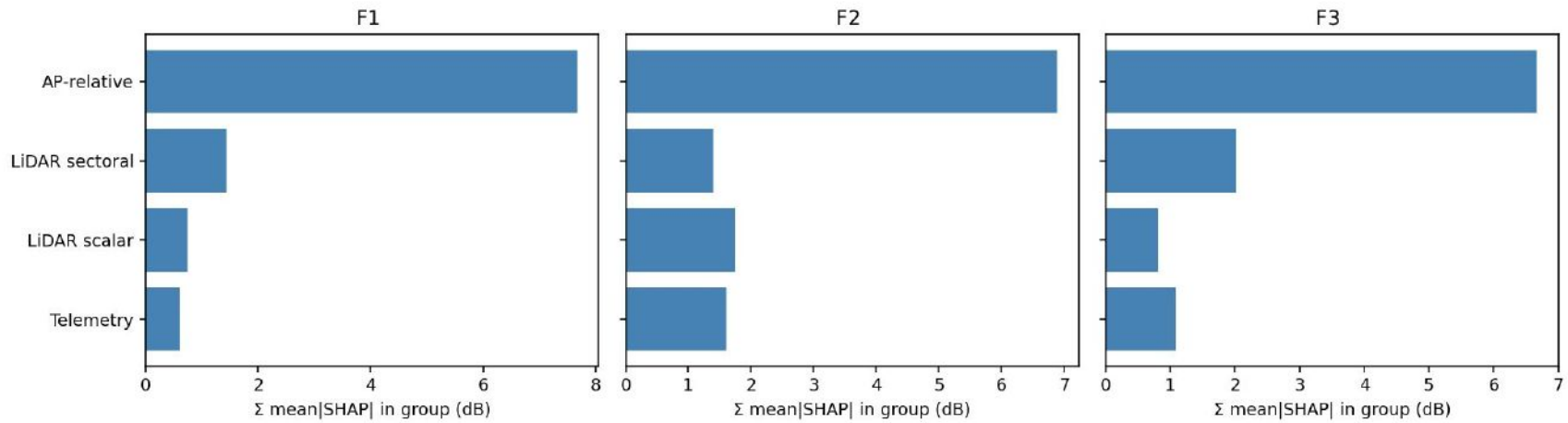
Action-grounded validation — let the measured effect of acting be the test. Promising but underformalised.

OPEN GAP Faithfulness has no ground truth — certifying “trustworthy enough to act” is open (Jacovi & Goldberg 2020).

Faithfulness as a recognised, still-open property: Jacovi & Goldberg 2020 (ACL).

CONDITION 1 · FAITHFULNESS — A STUDY WE CARRIED OUT

When attribution and attention disagree — a defensible negative



Cumulative attribution by feature group across three folds — the added sensor modality contributes least.

ρ = 0.078 Spearman rank correlation between attention and temporal attribution — essentially zero. Attention was not explanation here.

Attribution + a permutation-importance control gave a defensible negative: the simpler model (no added modality) sat within ~1 dB of the noise floor.

Discussion material from studies we carried out, restated generically. Temporal attribution: TimeSHAP (Bento et al. 2021).

CONDITION 2 · LOCUS

Thesis: one global feature-importance ranking suffices

THESIS

Rank the inputs once, globally. Read off the top features. Done.

One scalar per feature, averaged over the dataset — clean, and for many tabular problems, genuinely adequate.

CONDITION 2 · LOCUS

Antithesis: explanation has levels — a single ranking destroys structure

ANTITHESIS

- Explanation has levels — local vs global and beyond — a core question of the field (Guidotti et al. 2018; Barredo Arrieta et al. 2020).
- Static feature vectors destroy temporal structure (when) and modality structure (which sensor).
- Fusion XAI faces spatio-temporal misalignment across sensors and an intrinsic-vs-post-hoc trade-off (fusion + XAI survey, Sensors 2025).

"Which feature" is the wrong question when you need "when, from which sensor, for which decision."

Levels of explanation: Guidotti et al. 2018; Barredo Arrieta et al. 2020. Fusion XAI trade-offs: Sensors 2025, 25(3):856.

CONDITION 2 · LOCUS

Synthesis: a multi-level vocabulary — temporal · modality · decision

SYNTHESIS

(a)

Temporal attribution for when — e.g. TimeSHAP for recurrent models.

(b)

Modality attribution / ablation for which source — “what if a sensor were missing.”

(c)

Decision-level counterfactuals for the actual choices, like routing.

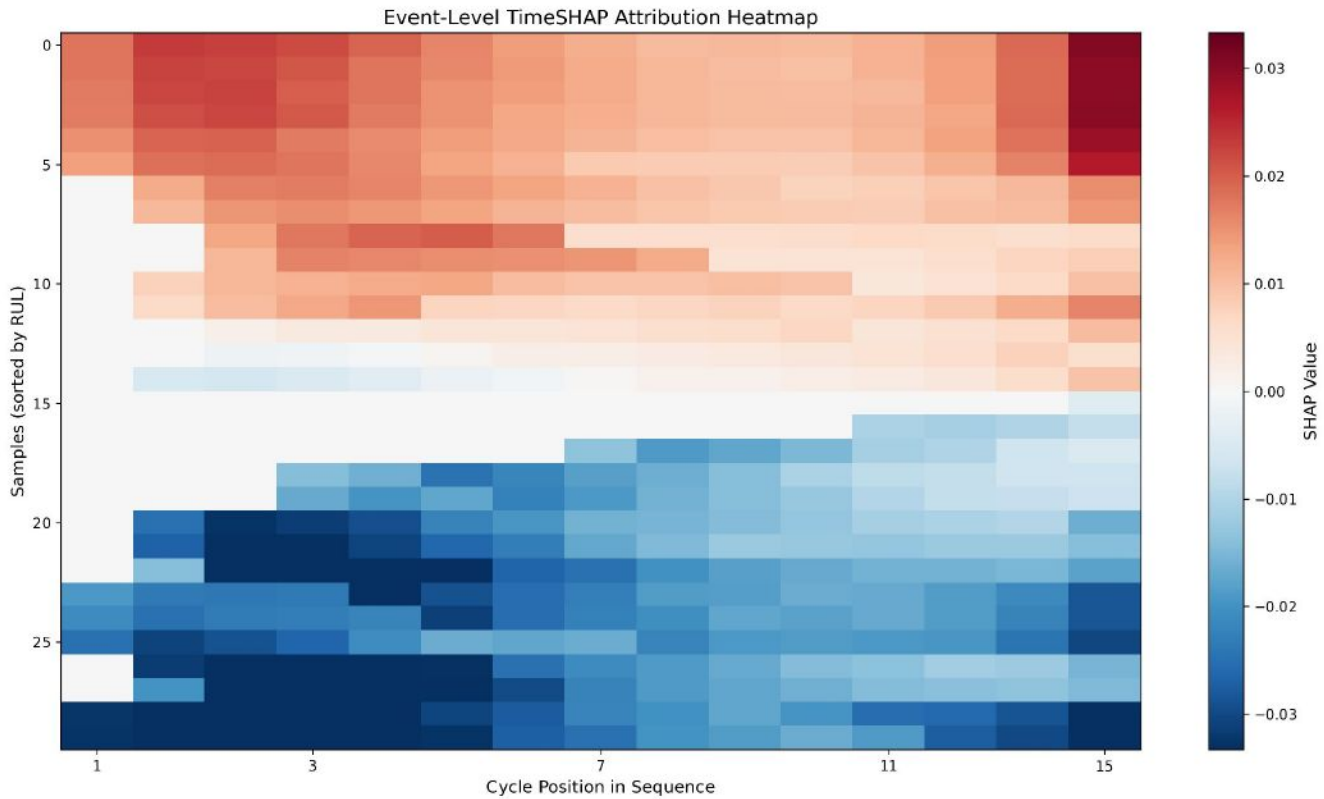
OPEN GAP

A unified, instance-level, environment-conditioned attribution spanning all three levels is still open.

Temporal attribution: TimeSHAP (Bento et al. 2021, KDD ’21). Intrinsic-vs-post-hoc trade-off: Sensors 2025, 25(3):856.

CONDITION 2 · LOCUS — A STUDY WE CARRIED OUT

Explaining “when”: temporal attribution on a recurrent prognostic model



Recent cycles dominate the prediction.

The “when” level, made visible: red raises the predicted remaining-useful-life, blue lowers it; the most recent steps carry the weight.

~16.9% of history can be pruned within 5% error — compute bought back on an edge device.

Related studies instantiate the other two levels: modality attribution (which sensor) and decision-level routing.

Discussion material, restated generically. Temporal attribution method: TimeSHAP (Bento et al. 2021).

CONDITION 3 · AGENCY

Thesis: a human stays in the loop

THESIS

Human-in-the-loop oversight and contestability — the regulator’s default (EU AI Act, Art. 14).

Morally and legally the obvious starting point: a person stays accountable.

Human oversight & contestability: EU AI Act, Art. 14.

CONDITION 3 · AGENCY

Antithesis: humans can’t track a real-time fusion loop

ANTITHESIS

- Humans over-rely and rubber-stamp — automation bias (EU AI Act Art. 14(4)(b); the NTSB cases).
- A real-time fusion loop, correcting many times a second, is too fast to supervise meaningfully.

→ **machine-in-the-loop self-adaptation** — monitor–analyse–plan–execute (MAPE-K, Kephart & Chess 2003), as in explanatory interactive learning.

Automation bias: EU AI Act Art. 14(4)(b); NTSB HWY18FH011 / HWY16FH018. Machine self-adaptation: MAPE-K (Kephart & Chess 2003); Ross et al. 2017.

Synthesis: a stakes × latency spectrum

SYNTHESIS

(a)

Bounded machine autonomy — re-weight / flag within a safety envelope; a human ratifies structural change.

(b)

Contestability — operators can always challenge and override.

(c)

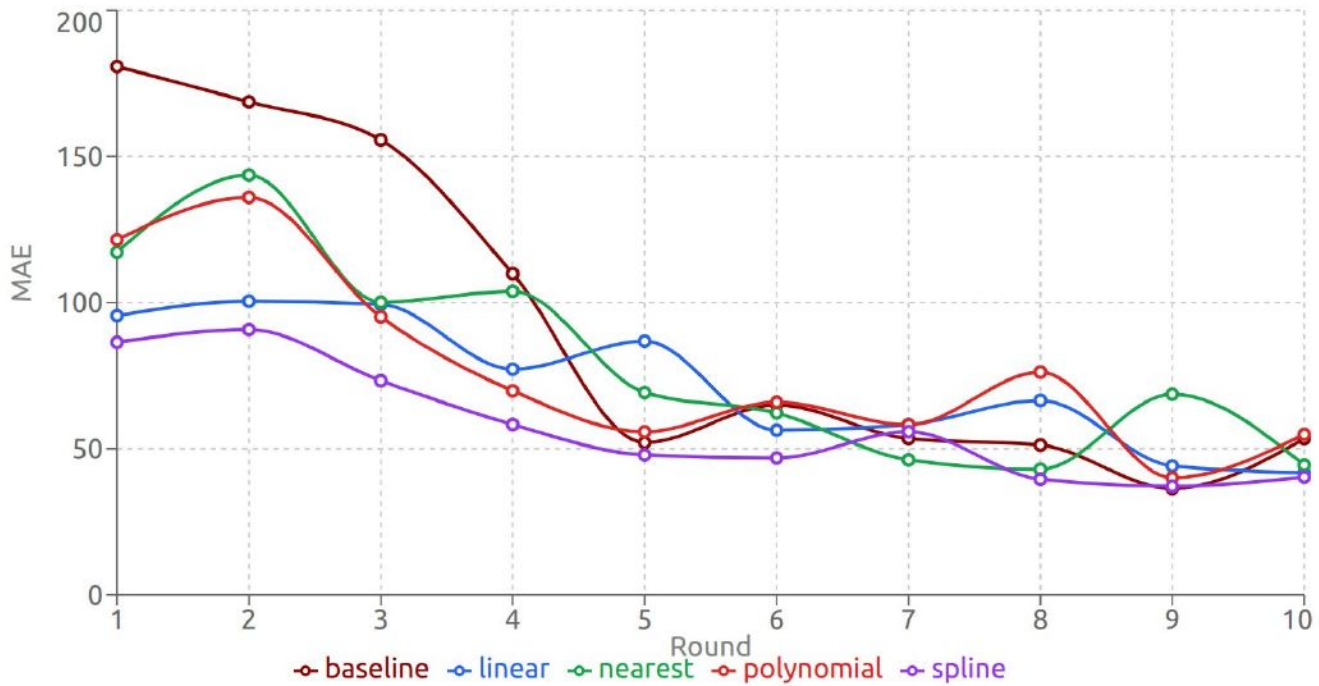
A triage policy — low-latency safety → machine; high-stakes structural → human.

OPEN GAP How to certify a partly self-modifying loop under Art. 14 is unresolved.

The regulation itself scales oversight to the "level of autonomy" (EU AI Act, Art. 14(3)).

CONDITION 3 · AGENCY — A STUDY WE CARRIED OUT

Both ends of the spectrum, in practice



Human end — decision support

Energy-aware routing informs the operator; the person decides.

Machine end — self-correction

Federated anomaly correction re-weights its own detectors; the volatile baseline (top) is tamed — ~25% error reduction.

A "detect → refine" loop needs a memory of what "correct" looks like — a requirement of its own.

Discussion material, restated generically. Federated anomaly correction across rounds; global forecasting error vs an uncorrected baseline.

Thesis: explainability is a free post-hoc layer

THESIS

Treat explanation as an analysis layer added after the fact. Compute is no object.

In a research notebook, on a workstation, that holds.

CONDITION 4 · REALITY

Antithesis: the loop has to survive contact with reality

ANTITHESIS

- A real-time-performance-vs-explainability trade-off, plus hard edge / energy limits (Sensors 2025).
- Loop stability — feedback can destabilise, oscillate, or be "gamed" (Maxwell 1868 lineage; specification-gaming, Amodei et al. 2016).
- Explanations are mostly used to debug models, not inform users — thin deployment evidence (Bhatt et al. 2020).

Reality is where elegant loops go to die.

Real-time vs explainability & edge limits: Sensors 2025, 25(3):856. Stability/gaming: Maxwell 1868; Amodei et al. 2016. Deployment reality: Bhatt et al. 2020.

CONDITION 4 · REALITY

Synthesis: lightweight · stable · measurable

SYNTHESIS

(a)

Lightweight / edge XAI — amortised or approximate attribution, distillation, on-device pruning.

(b)

Control-theoretic safeguards — rate limits, hysteresis, damping against oscillation / gaming.

(c)

Validation — reproducible benchmarking + the loop's measured reliability gain.

OPEN GAP No standard benchmark for "explanation-driven control improved the system."

Stability safeguards from control theory; validation via reproducible benchmarking + measured reliability gain.

CONDITION 4 · REALITY — STUDIES WE CARRIED OUT

Lightweight, measurable, and reproducible — in practice

~25%

forecasting-error reduction from the federated correction loop.

~17%

of history prunable within 5% error — the explanation buys back edge compute.

Open

simulation benchmark isolating sensor-quality effects — reproducibility.

Green + human

energy-aware routing; worker physiological-load (macro-F1 ≈ 90.7%, HR error < 10 bpm).

None of these is the whole loop — but each is a real, measured piece of it.

Discussion material, restated generically. Worker-load study: macro-F1 ≈ 90.7%, HR MAE < 10 bpm. Open simulation benchmark for sensor-quality effects.

AN OPEN RESEARCH AGENDA

Four open problems — one per condition

Faithfulness	Validation without ground truth — can action-grounded validation be made rigorous?
Locus	A unified, instance-level, environment-conditioned fusion explanation.
Agency	Certifying a partly self-modifying loop.
Reality	A shared benchmark for “explanation-driven control improved the system.”

And the cross-cutting thread: a reference “memory of what correct looks like.”

Selected references

PRECEDENTS & CONTROL LINEAGE

Maxwell, J. C. (1868). On Governors. Proc. R. Soc. Lond. 16:270–283.

Shewhart, W. A. (1931). Economic Control of Quality of Manufactured Product. Van Nostrand. (First chart: Bell Labs memo, 16 May 1924.)

Medtronic MiniMed 670G — first hybrid closed-loop system; US FDA approval, 28 Sep 2016.

Åström, K. J. & Murray, R. M. (2008). Feedback Systems. Princeton Univ. Press.

Fikes, R. & Nilsson, N. (1971). STRIPS. Artificial Intelligence 2(3–4):189–208.

Kephart, J. & Chess, D. (2003). The Vision of Autonomic Computing (MAPE-K). IEEE Computer 36(1):41–50.

EXPLANATION-AS-CONTROL

Ross, A., Hughes, M. & Doshi-Velez, F. (2017). Right for the Right Reasons. IJCAI, 2662–2670.

Teso, S. & Kersting, K. (2019). Explanatory Interactive ML. AIES, 239–245.

Schramowski, P. et al. (2020). Nat. Mach. Intell. 2(8):476–486.

Amodei, D. et al. (2016). Concrete Problems in AI Safety. arXiv:1606.06565.

XAI METHODS & DEBATE

Ribeiro, M. T. et al. (2016). LIME. KDD ’16, 1135–1144.

Lundberg, S. & Lee, S.-I. (2017). SHAP. NIPS 2017, 4765–4774.

Bento, J. et al. (2021). TimeSHAP. KDD ’21, 2565–2573.

Rudin, C. (2019). Nat. Mach. Intell. 1:206–215.

Jain, S. & Wallace, B. (2019). Attention is not Explanation. NAACL-HLT, 3543–3556.

Wiegreffe, S. & Pinter, Y. (2019). Attention is not not Explanation. EMNLP-IJCNLP, 11–20.

Krishna, S. et al. (2022). The Disagreement Problem. arXiv:2202.01602.

Slack, D. et al. (2020). Fooling LIME and SHAP. AIES, 180–186.

Jacovi, A. & Goldberg, Y. (2020). ACL, 4198–4205.

Guidotti, R. et al. (2018). ACM Comput. Surv. 51(5):93.

Barredo Arrieta, A. et al. (2020). Information Fusion 58:82–115.

Bhatt, U. et al. (2020). Explainable ML in Deployment. FAccT, 648–657.

Domain / regulation / stats: EU AI Act Arts. 13 & 14 (14(4)(b), automation bias; high-risk via Art. 6 + Annex I). Yeong, Panduru & Walsh (2025), Sensors 25(3):856. NSC Injury Facts — Forklifts (BLS CFOI/SOII). Senseye/Siemens, True Cost of Downtime 2024. McKinsey, “Prediction at scale.” NTSB HWY18FH011 (2020); HWY16FH018.

Instrument set: artifact → metric → test

One instrument set, run under every background: RQ1 = metric values under B0/B1; RQ2 = deltas under candidates (M1–M6, one per artifact A1–A6).

Example — A1 edge inflation / M1: Spearman correlation between attribution magnitude and substitution distance on known-irrelevant positions (Design-3 trap channels); 0 is ideal.

Every row carries a competence gate (model must exhibit the prerequisite behaviour before an effect is attributed to the artifact) and a confound control (irrelevant positions defined by construction, not by the explainer).

TAB — admissibility and definitions

R1 — past-only measurability: b(t; x) depends on $x_1..$□ and label-free training statistics only. Normalized position t/T is inadmissible (T encodes remaining life); elapsed age t is admissible.

R2 — coalition-independence: the substitute is computed once from the observed sequence — one precomputation serves every coalition (the near-free cost claim).

R3 — no shape assumption: no monotonicity, linearity, or parametric trend form; shape constraints only as declared per-profile variants.

TAB-state (B2): empirical mean of training observations in the same observable-state cell (quantile-binned past-only state index).
TAB-traj (B3): strictly-past smooth fit of the sequence's own level, evaluated at the masked position — masking removes the timestep's innovation. Fallback to B2 below a history threshold.

Benchmark pairing

Profile	Paired real dataset	Property under test
P-PCoE	NASA PCoE	knee-shaped trajectories; documented counter-trend excursions; case-study continuity
P-CALCE	CALCE CS2/CX2	replication under a second lab protocol and cell format
P-SEV	Severson / MATR	many sequences, flat early trajectory → statistical power; A2 stress
P-RW	NASA RW	rough trajectory, frequent load-driven transients → A3/A4 stress
P-GEN	— (synthetic-only)	non-monotone mean-nonstationarity → mechanism-level scope probe
P-CMAPSS (opt.)	C-MAPSS	different physical domain and trajectory morphology

Adaptive replay — signals and metrics

DEPLOYMENT SIGNALS (no ground truth)

explanation outputs · explanation confidence / stability · model confidence · sensor reliability estimates · disagreement between sensors · temporal consistency · delayed task outcomes (if available)

EVALUATION METRICS

task accuracy · collision rate · stopping distance · robustness · calibration · adaptation speed · recovery after degradation · replay efficiency