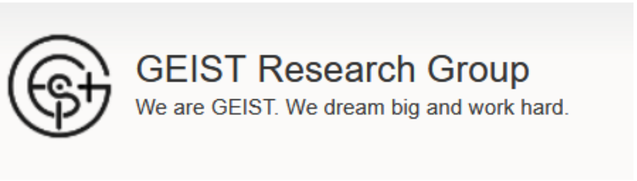
Transparent and Adaptive AI for Human-Guided Decision Support.

PhD Candidate: Sabri Manai
Supervised by: Grzegorz J. Nalepa and Szymon Bobek



We see that we are muted on Teams.
We still whisper.

So how can we expect people to trust AI systems
with decisions that actually matter?

The problem is not only black-box AI

People do not just need to know why a model made a prediction.
They need to work with systems that can handle questions like:

What changed?

Can the system learn from my feedback?

What are my options?

What happens if I choose differently?

This shifts the problem from explaining a single output
to supporting an ongoing decision process.

Hypothesis & research direction

Can combining explainability, adaptability, and interactivity make AI systems more trustworthy and useful than black-box approaches?

Research Trajectory

Work 1 → Explanations for ditterent stakeholders.
(Who needs to understand?)

Work 2 → Adding domain context.
(What background knowledge helps?)

Work 3 → Actionable countefactual guidance.
(What should the user do next?)

Current → LLM-based interactive explanation interface.
(Does it work for real operators in real settings?)

Research Trajectory

Work 1 → Explanations for ditterent stakeholders.
(Who needs to understand?)

Work 2 → Adding domain context.
(What background knowledge helps?)

Work 3 → Actionable countefactual guidance.
(What should the user do next?)

Current → LLM-based interactive explanation interface.
(Does it work for real operators in real settings?)

Research Trajectory

Work 1 → Explanations for ditterent stakeholders.
(Who needs to understand?)

Work 2 → Adding domain context.
(What background knowledge helps?)

Work 3 → Actionable countefactual guidance.
(What should the user do next?)

Current → LLM-based interactive explanation interface.
(Does it work for real operators in real settings?)

Research Trajectory

Work 1 → Explanations for ditterent stakeholders.
(Who needs to understand?)

Work 2 → Adding domain context.
(What background knowledge helps?)

Work 3 → Actionable countefactual guidance.
(What should the user do next?)

Current → LLM-based interactive explanation interface.
(Does it work for real operators in real settings?)

Research Trajectory

Work 1 → Explanations for ditterent stakeholders.
(Who needs to understand?)

Work 2 → Adding domain context.
(What background knowledge helps?)

Work 3 → Actionable countefactual guidance.
(What should the user do next?)

Current → LLM-based interactive explanation interface.
(Does it work for real operators in real settings?)

Work 1

Explainadble Next-Purchase Recommendations

- Multistakeholder explanations — tailored for end-users and business stakeholders
- SHAP-based, genre-level attributions for end-users
- Latent-factor interpretation for business/marketing stakeholders
- Serendipity & cold-start support (novelty balanced with business rules)

Explanations adapted to different users and decision needs.

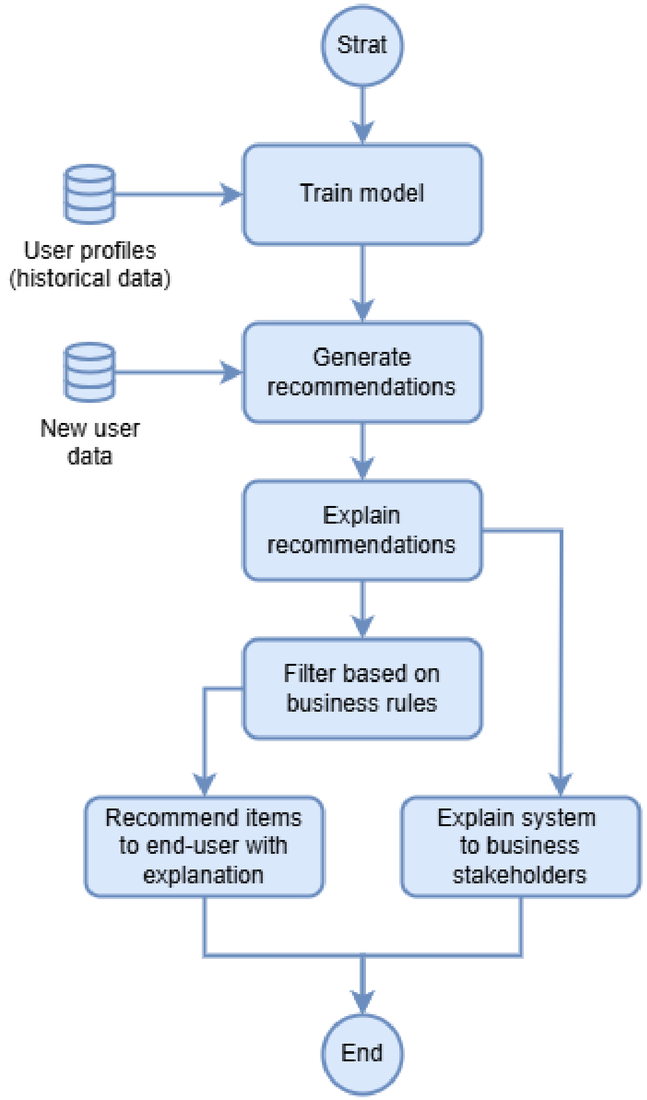


Fig1: High-level architecture of , multistakeholder recommendation system.

Source: Mozolewski et al., Explainable Next-Purchase Recommendations: A Multistakeholder Framework, 2025.

Work 1

Linking latent factor to movie genre

- Linking latent factors to geners
- Some geners strongly corrolate with more than one latent factor
- We used this as a base for explanations for stakeholders

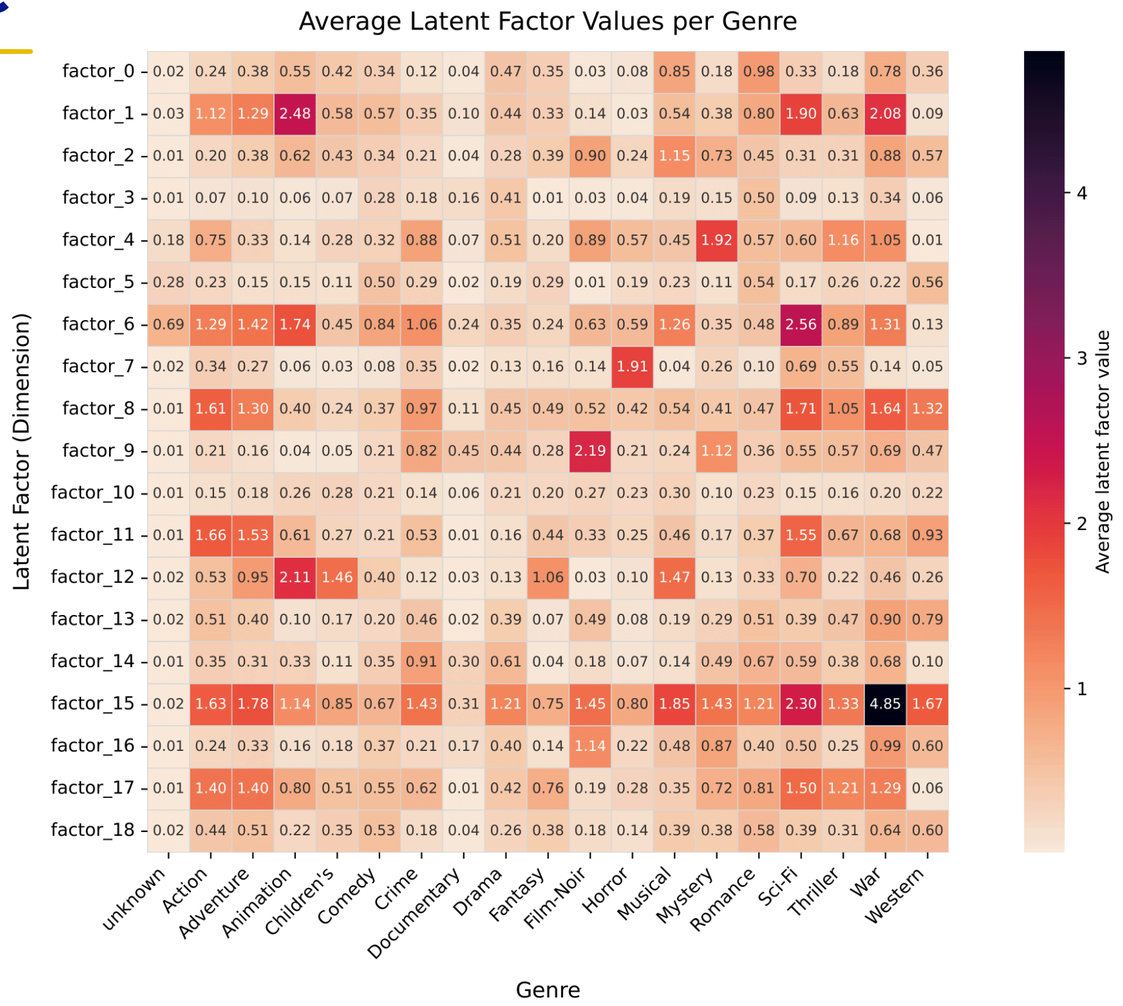


Fig 2: Average latent-factor strength per genre. Darker cells mark stronger associations

Source: Mozolewski et al., Explainable Next-Purchase Recommendations: A Multistakeholder Framework, 2025.

Work 2

Knowlendge Graphs for Explainable Industrial AI

- Structured domain knowledge connects raw signals to semantic context
- Anomalies interpreted in context, not as isolated alerts
- Relevant to anomaly detection, predictive maintenance, root-cause analysis
- Knowledge-driven dashboards that build operator trust

Work 3

Counterfactual User Guidance for Improving Transparent Hyperparameter Tuning

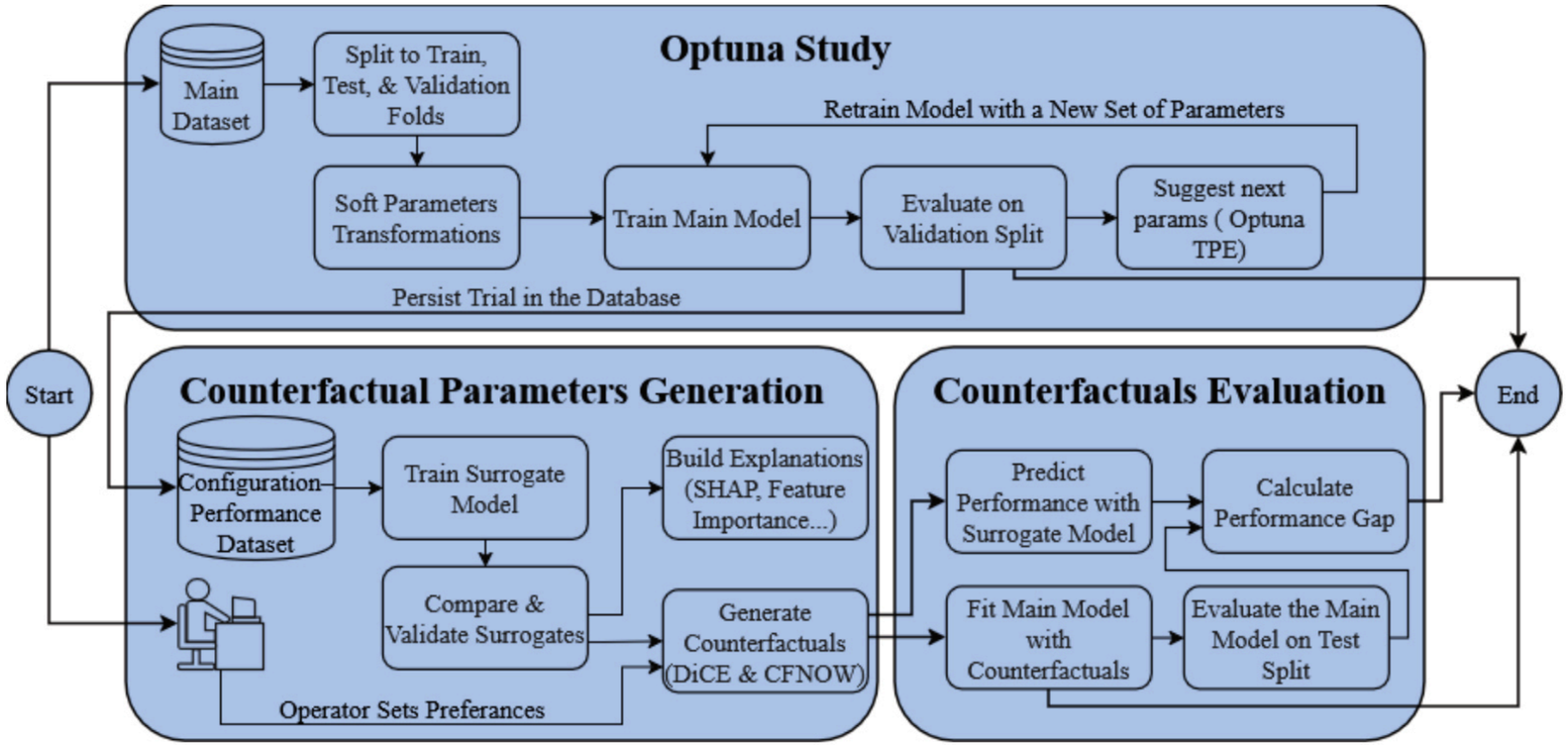


Fig 3: End-to-end pipeline for counterfactual-based hyperparameter optimization.

Source: Manai et al., Counterfactual-guided Explainable Hyperparameter Optimisation, ICCS 2026, in press.

Generally couneterfactuals are faithful to the surrogate model

- DiCE generally produces counterfactuals that stay close to the surrogate estimates
- DiCE often yield small but consistent gains when the model is retrained.
- CFNOW generates more conservative counterfactuals, which in some datasets do not translate into improvements for the main model.

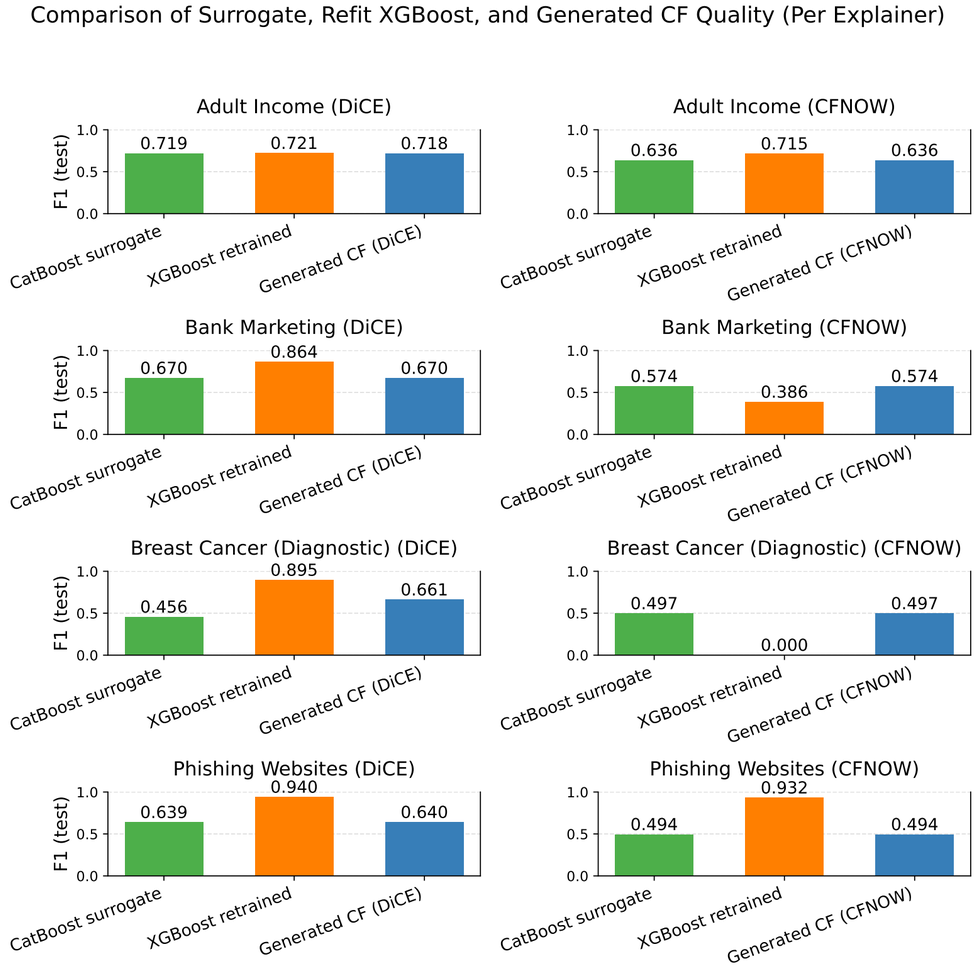


Fig 4: Counterfactual quality across datasets for DiCE and CFNOW

Source: Manai et al., Counterfactual-guided Explainable Hyperparameter Optimisation, ICCS 2026, in press.

Current Work

Journal extension & user study

- Extending counterfactual HPO toward anomaly detection
- Building and evaluating an interactive interface
- Comparing interface-based guidance and a notebook baseline
- Testing whether users actually understand constraints, trade-offs, alternatives

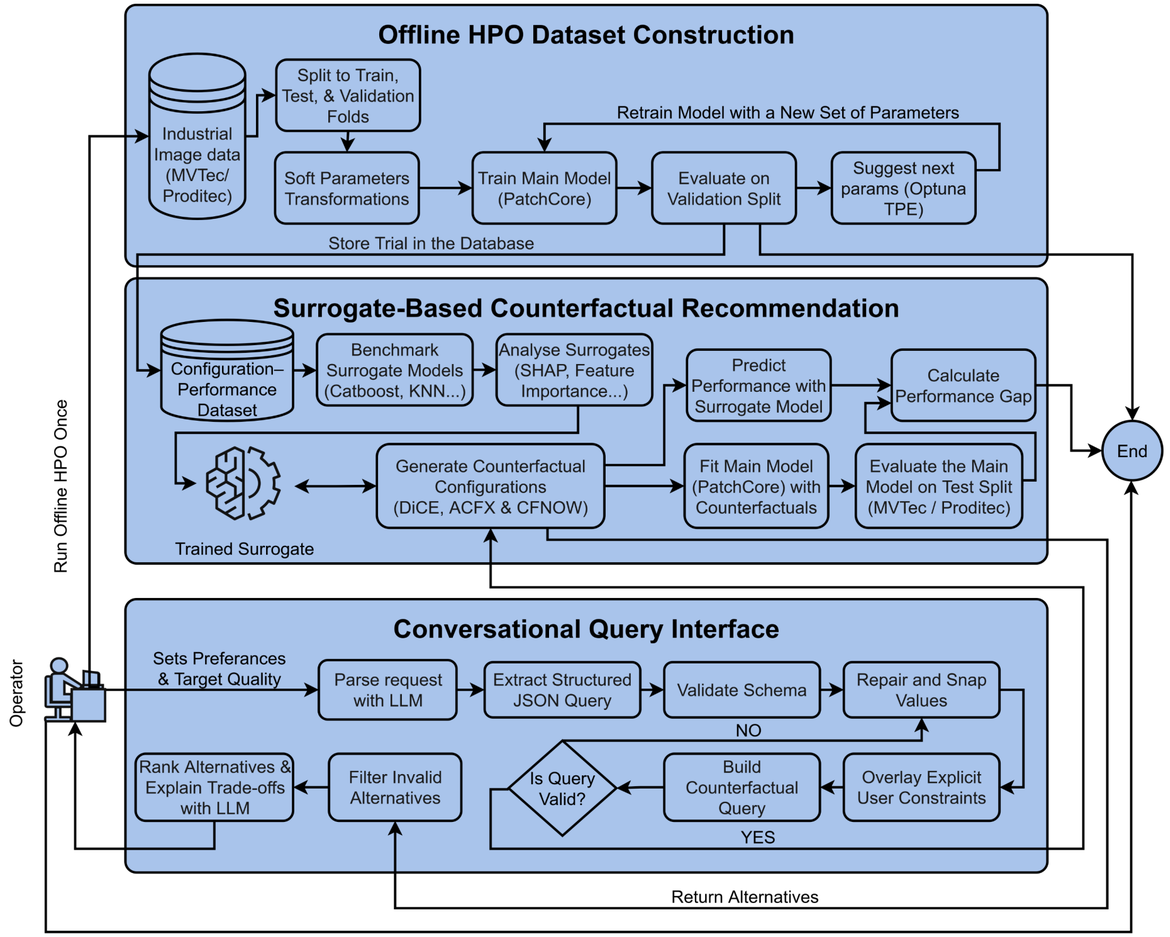


Fig 5: Conversational Counterfactual Guidance for Hyperparameter Optimisation

The conversational-interface condition had higher mean scores on all five items.

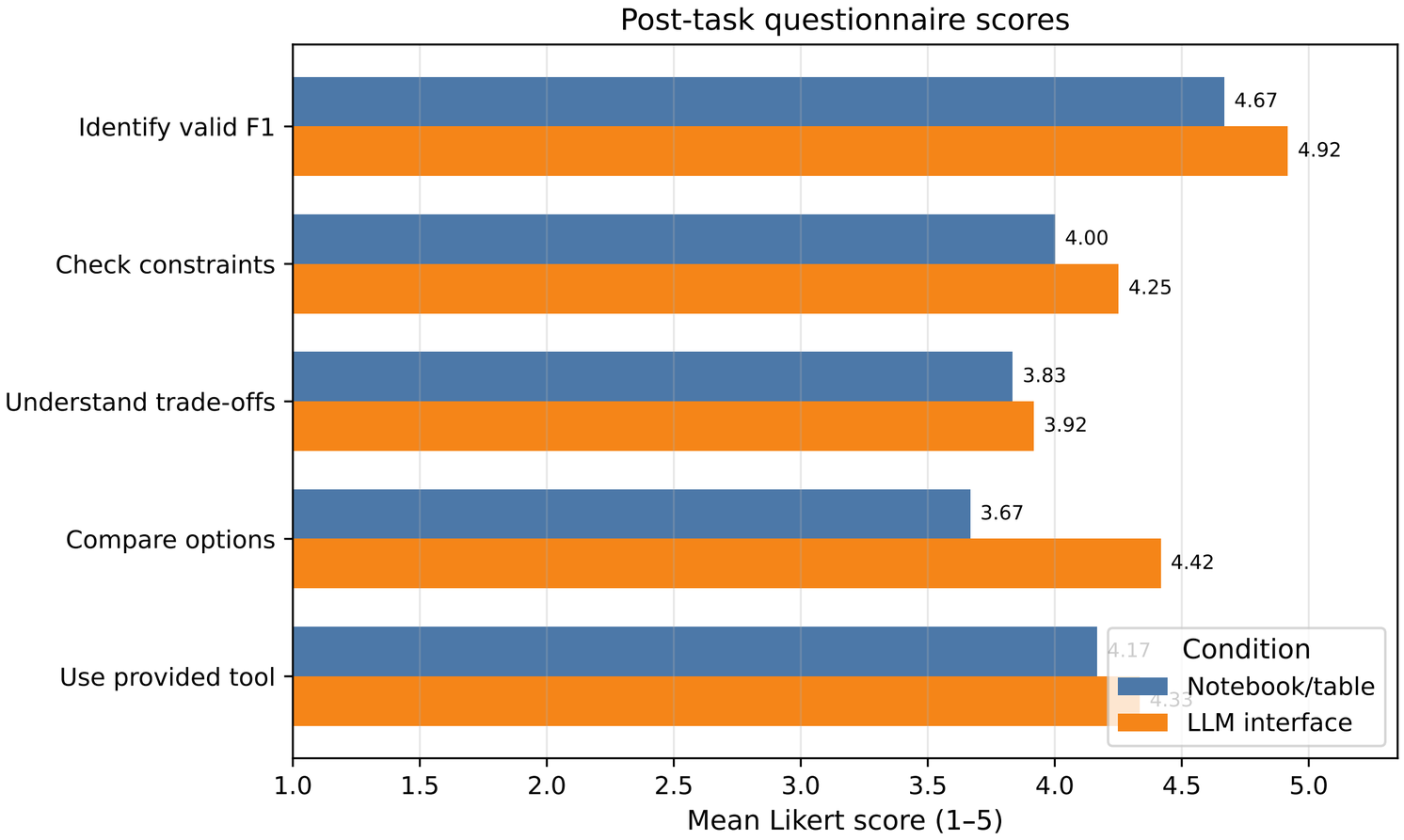


Fig 6: Post-task questionnaire scores by condition. Scores are mapped Lik-

The largest difference was for comparing options (LLM interface: 4.42 vs. notebook/table: 3.67)

These differences should be treated as indicative of perceived usability differences rather than as reliable effect estimates.

Next Phase

From one-shot explanations to adaptive interaction

- Adaptive conversational interfaces for pattern understanding
- Feedback loops to validate, correct, and refine outputs
- Explanation-aware interaction, not static explanations
- Systems that adapt to preferences, corrections, and drifting data
- Evaluating effects on trust, understanding, and decision quality
- Extension to multimodal settings where relevant

The real open problem: disagreement

When a user pushes back on an AI explanation, current systems do one of two things:

- **Ignore it**: rigid, unresponsive
- **Accept it:** sycophantic, unprincipled

Neither is right. The honest response requires the system to explain its own reasoning about the disagreement.

When should the system hold, when should it revise,
and how should it communicate that choice?

One-shot explanations are not enough

Three things happen after an AI makes a decision:

1. The user has questions the explanation doesn't answer
2. The user disagrees with the reasoning
3. The context changes and the decision needs revisiting

A single explanation handles none of these.
It's a snapshot, not a conversation.

The evidence confirms it

- Users shown AI explanations grow more confident,
  even when the AI is wrong

- When users stop over-relying on AI, they don't automatically
  make better decisions they just disagree more

The evidence confirms it

- Users shown AI explanations grow more confident, even when the AI is wrong
- When users stop over-relying on AI, they don't automatically make better decisions they just disagree more

Richer explanation shifts how users *feel*.
Not necessarily what they decide.

What the field has been building

The last two years have produced real progress:

- Interactive interfaces that let users choose which explanation to see *(IXAII, 2025)*
- Adaptive XAI workshops *(IUI 2024, 2025)*
- *Conversational XAI*: LLM-based dialogue where users ask follow-up questions and request clarification

This is the right direction, but most of the systems still assume that the explanation adapts to the user.

The loop is still one-directional

What exists	What's missing
Choose your explanation type	Respond when user contests the reasoning
Adapt explanation to user context	Adapt the decision based on feedback
Ask follow-up questions	Hold, revise, or surface disagreement

The closest works and what is left open

Belief-change theory for XAI (Coba et al., 2024): Formalizes how feedback should revise explanations logically. Principled and rigorous. No system built, no users tested.

IXAII (Speckmann et al., 2025): Gives users agency over which explanation to view. The system doesn't respond to disagreement with its reasoning.

Contestability in AI (Pi et al., FAccT 2026): Studies how people contest AI decisions at governance and societal level. Not about the system responding in real time to a single user.

When a user contests an AI explanation, when should the system hold, revise, or surface the disagreement, and does getting this right produce better human decisions?

The research question

Three sub-questions, three chapters

1.

What happens when users can contest an explanation?

2.

How should a system decide to hold, revise, or surface disagreement?

3.

Does principled contestation produce better decisions and appropriate reliance?

Assumptions Stated Upfront

Principled adaptation will improve appropriate reliance over static explanation but not uniformly.

It will fail when:

- The system is confidently wrong and holds anyway
- The user is an expert and the system overrides a correct correction
- The interaction loop increases confidence without improving accuracy

Step 1:
What happens when users can contest an explanation?

The literature gives users a static explanation and measures what they do with it. The moment of disagreement, when the user thinks the system is wrong, is treated as noise, not data.

What we need to observe?

When a user contests an explanation, these three things matter:

- **What they contest:**
  the conclusion, the reasoning, or the evidence behind it?
- **How they contest it:**
  challenge, propose an alternative, ask for justification?
- **What they expect:**
  the system to explain itself, adapt, or acknowledge the conflict?

How do we approach it?

The goal is to observe and understand contestation
before designing for it.

The approach:
Present users with AI recommendations and explanations, including flawed ones, and observe what they question, how they express disagreement, and what they expect from the system in response.

The output is a structured understanding of contestation
that step 2's mechanism is built on.

Step 2
When should the system hold, revise, or surface disagreement?

Step 1 maps what users contest and how.
Step 2 asks what should the system do with that?

The field has the formal foundation, belief-change theory tells us revision should be principled. What doesn't exist is a working mechanism that operationalizes this in a real decision-support system.

Three possible responses to contestation

When a user pushes back, the system has three honest options:

Hold

The evidence supports the original recommendation; the system explains why.

Revise

The user's input is consistent with the evidence; the system updates and explains what changed.

Surface

The evidence is genuinely uncertain; the system exposes the conflict rather than resolving it silently.

What drives the decision

The hold/revise/surface choice is not based on user preference or social pressure. It is grounded in:

- How confident the system is in its recommendation
- Whether the user's alternative is consistent with the underlying evidence
- How uncertain the model is in the region the user is contesting

The explanation of the response

Choosing how to act is not enough.
The system must explain why it responded that way.

> *"I'm holding because the evidence strongly supports this recommendation"*

> *"I'm revising because your input is consistent with what the data shows"*

> *"I'm surfacing a conflict because the evidence here is genuinely uncertain"*

What are the outcomes of this step?

- A concrete, evidence-grounded mechanism for decision-support systems
- An operationalization of belief-change theory in a working system
- The building block for step 3 to evaluate against human decision quality

Step 3

Does principled contestation actually improve human decisions?

Step 2 defines what the system should do.
Step 3 asks whether it makes a difference to the human on the other side.

The measure is not only perceived trust or satisfaction.
It is whether users make better decisions , and whether they rely on the system appropriately, not just confidently.

Appropriate reliance — the right measure

Most XAI evaluations measure trust. Trust is not enough.

A user can trust a system that is wrong.
A user can distrust a system that is right.

Appropriate reliance:
The user accepts the system's recommendation
when it is correct, and overrides it when it is not.

This is the measure that connects explanation quality
to decision quality.

Study design

Three conditions:

- **Static explanation** — one-shot, no interaction
- **Unconstrained interaction** — user can push back, system adapts freely
- **Principled contestation** — the hold/revise/surface mechanism from Step 2

The question:

Does principled contestation produce better appropriate reliance than both baselines, and under what conditions does it fail?

Pre-committed failure conditions

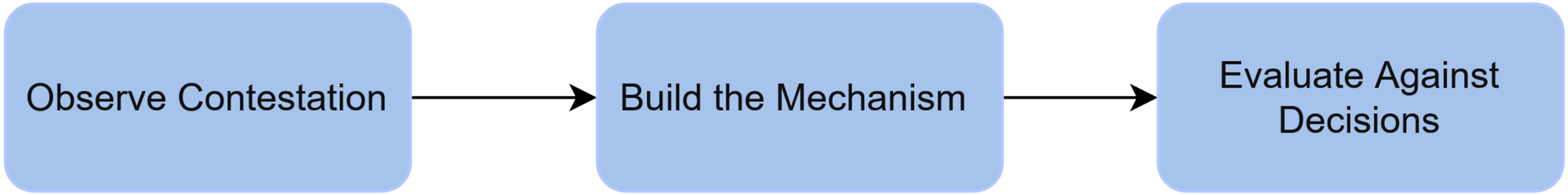
The mechanism is expected to fail:

- The system holds confidently but is wrong
- The user is right and the system overrides a correct correction
- Interaction increases confidence without improving accuracy

What are the outcomes of this step?

- The first evaluation of this mechanism against appropriate reliance as the target measure
- A characterization of when principled contestation helps and when it doesn't
- An answer to the thesis question, conditional, honest, and empirically grounded

Summing up future directions



Step 1 grounds the mechanism in observed human behavior.

Step 2 builds what the field has formally described but not yet implemented.

Step 3 tests whether it really matters.

Two honest risks and what limits them:

Mechanism underspecified:
Translating hold/revise/surface into a working system requires concrete design decisions that are not yet made.

Recruitment and study validity:
A powered, behavioral study with the right population takes time and access. Testing with real operators in real settings is hard to control and harder to schedule.

What this thesis contributes

1. Explainability for multiple stakeholders
2. Domain-grounded industrial AI connecting model outputs to semantic context through knowledge graphs
3. Counterfactual-guided interaction showing that conversational explanation outperforms static alternatives
4. A principled contestation mechanism
5. An empirical account of appropriate reliance meaning when interactive explanation helps human decisions.

Thank you!

Email: **sabri.manai@uj.edu.pl**

- *Manai, S., Bobek, S., Nalepa, G.J., do Valle Miranda, L., Kutt, K., Jung, J.J. (2026). Toward Explainable Industrial AI: The Role of Knowledge Graphs. In: Martínez, L., et al. Intelligent Data Engineering and Automated*
- *Mozolewski, Maciej & Zych, Honorata & Manai, Sabri & Kutt, Krzysztof & Nalepa, Grzegorz. (2025). Explainable Next-Purchase Recommendations: A Multistakeholder Framework.*
- *Manai, S., Bobek, S., Nalepa, G.J, Counterfactual-guided explainable hyperparameter optimisation, in: Proceedings of the International Conference on Computational Sci-*