Visible AI

Does Seeing the Model Help Users Learn?

Agenda

01.
Introduction

02.
Theoretical Foundation

03.
Methods

04.
Expected Contributions

05.
Discussion

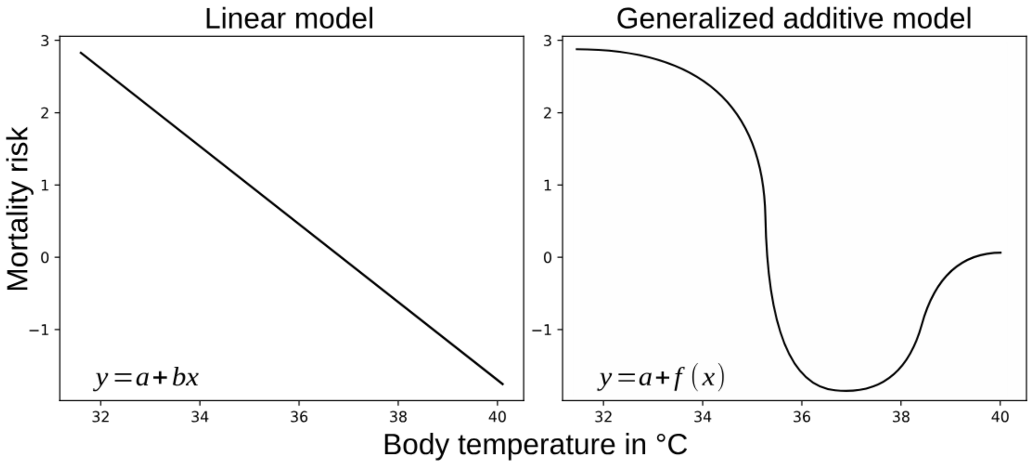
01 Introduction

Intrinsically Interpretable Models: LMs and GAMs

- GAMs represent the **relationship between input features and the target variable** using so-called **shape functions**
- The final prediction is obtained by summing the contributions of each shape function:

$$f(x) = f_1(x_1) + f_2(x_2) + \cdots + f_n(x_n)$$

- **Example:** Modeling mortality risk based on body temperature



The Core Research Question

When people work with AI systems repeatedly, does seeing how the model works help them learn better than just seeing its predictions? *And what types of learning do different explanations support?*

Why This Matters

- Most AI interaction research focuses on single decisions
- **Real-world reality**: People work with AI systems over time and develop two types of expertise: **Domain expertise** (understanding actual task relationships) and **AI expertise** (understanding AI behavior)
- **Current gap**: We don't understand how explanation transparency affects learning trajectories
- **Practical need**: When should we show explanations for **domain learning** vs. **AI collaboration**?

The Learning vs. Understanding Distinction

Current Research Focus: Static Understanding

- "Can you interpret this explanation?"
- "Do you trust this prediction?"
- "How satisfied are you with this interface?”

Our Focus: Dynamic Dual Learning

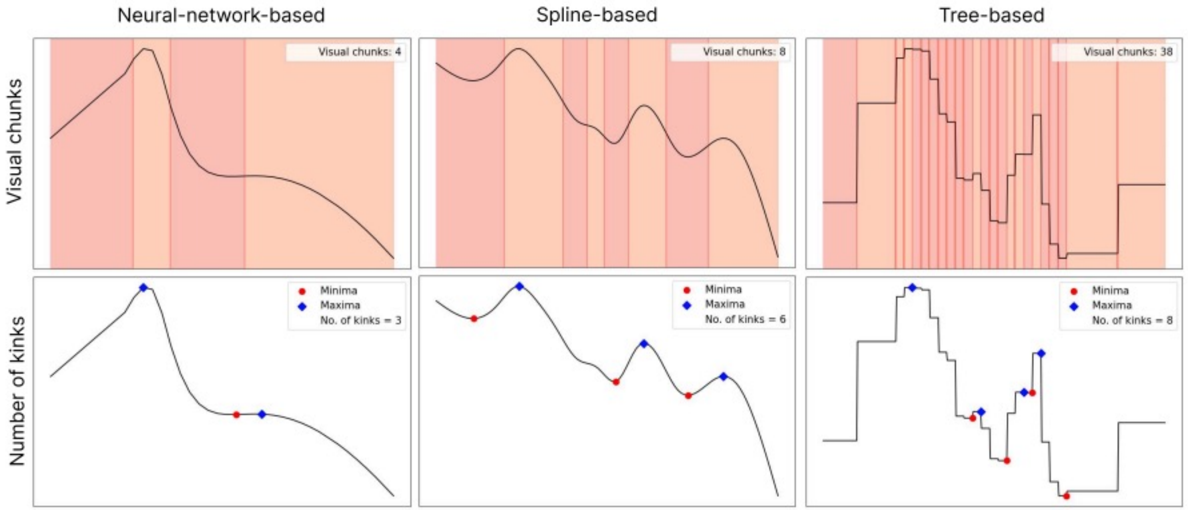
- "Do you get better at predicting **ground truth** over time?"
- "Do you learn to predict **what the AI will say**?"
- "Do you develop **domain expertise** vs. **AI expertise**?"

Understanding an explanation ≠ Learning from repeated interaction ≠ **Learning the right things**

02 Theoretical Foundation

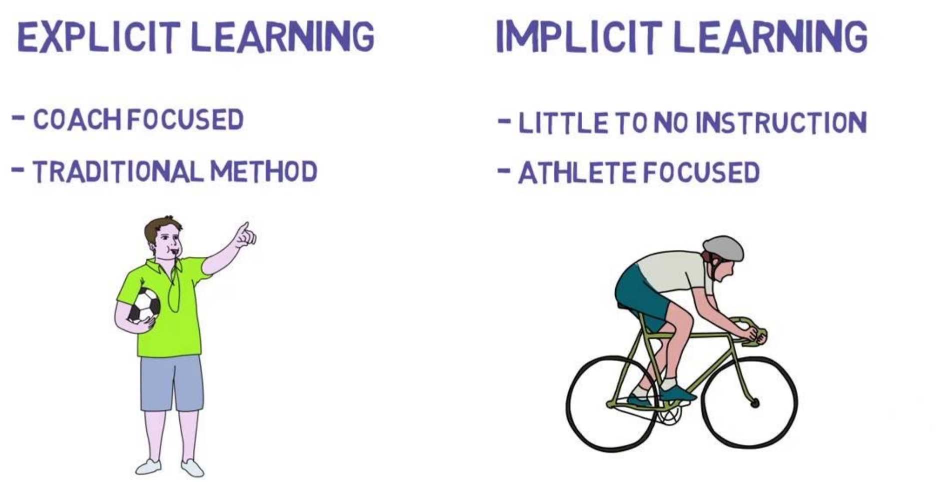
Building on Cognitive Load Research

- Kruschel et al. (2024): Visual complexity in GAMs affects cognitive processing
- **Our extension**: How does this affect learning over multiple interactions?
- **Complexity manipulation**: Systematic variation in GAM curve complexity



Learning Theory Integration

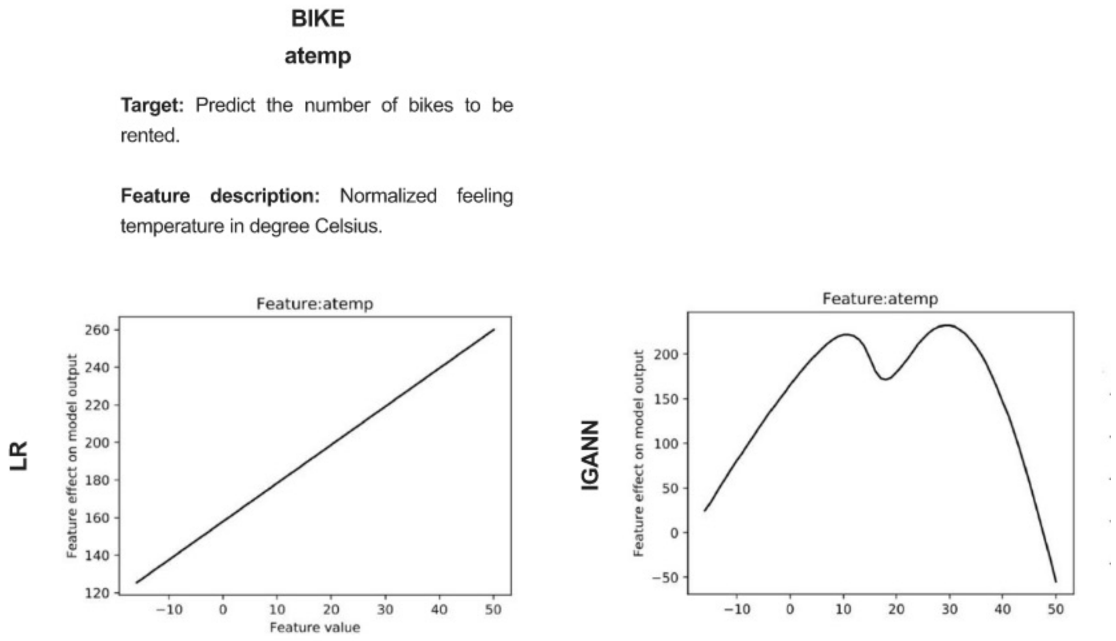
- **Explicit learning**: Direct instruction from visible explanations
- **Implicit learning**: Pattern recognition from prediction observation
- **Individual differences**: Who learns better from which approach?



03 Methods

Domain: Bike Rental Prediction

- Intuitive domain with learnable patterns
- Objective performance measurement (prediction accuracy)
- Real-world AI application (demand forecasting)



Treatment Conditions: Model Visibility

Visible Conditions

- **Visible Linear Model**: See straight-line plots showing how features affect rentals
- **Visible GAM**: See curved plots showing complex, non-linear relationships

Non-Visible Conditions

- **Non-Visible Linear**: Get AI predictions but no explanation of how it works
- **Non-Visible GAM**: Get AI predictions but no explanation of how it works

Control

- **No Model**: Just your predictions vs. truth (pure human learning)

Core question: Do explanations help **domain learning** vs. **AI behavior learning**, or can people figure out both types of patterns just from predictions?

What We Measure: Learning Over Time

Two Types of Learning:

- **Domain Learning**: How well do you predict actual bike rentals?
- **AI Learning**: How well do you predict what the AI model will say?

Key Questions:

- Which explanation types support which kind of learning?
- Do people become domain experts or AI experts?
- Who transfers better to independent work?

Research Design: A Learning Game

Phase 1: Learning Phase (8-10 rounds)

- **Your prediction first**: Based on weather/date/... features
- **AI prediction revealed**: You see what the model predicts (+ explanation depending on condition)
- **Final prediction**: You can adjust based on AI input
- **Truth revealed**: See actual bike rentals + your error vs. AI error
- **Payment**: Bonus for accuracy + correctly identifying when AI is wrong

Phase 2: Knowledge Assessment

- Closed-book questions about relationships
- "When is temperature most important for bike rentals?"
- "What weather patterns make the AI unreliable?”

Phase 3: Application Phase (5-6 rounds)

- No AI assistance - apply what you learned
- Predict both: actual rentals AND what the AI would predict
- Test: Did you truly internalize the patterns?

Variables

Independent Variables (X)

Primary Factor: Model Visibility (Between-subjects)

- **Visible Linear Model**: See straight-line plots explaining AI decisions
- **Visible GAM**: See curved plots showing complex relationships
- **Non-Visible Linear Model**: Get AI predictions with no explanation
- **Non-Visible GAM**: Get AI predictions with no explanation
- **No Model Control**: Pure human learning from feedback

Individual Difference Moderators

- Spatial reasoning, numeracy, graph literacy, domain knowledge

Dependent Variables (Y)

Primary Outcomes:

- **Learning Rate**: Improvement in prediction accuracy across 8-10 rounds
- **Knowledge Internalization**: Performance on closed-book relationship questions
- **Transfer Performance**: Accuracy in application phase without AI assistance

Secondary Outcomes:

- **Error Detection**: Ability to identify when AI predictions are wrong
- **Confidence Calibration**: Knowing when you know vs. don't know

Key Hypotheses: Dual Learning

- **H1:** Visible conditions will show faster learning than non-visible conditions, but the advantage will differ by learning type
- **H2:** Visible GAM will best support domain learning (ground truth prediction); Visible Linear will best support AI behavior learning (model prediction)
- **H3:** Domain expertise (from GAM explanations) will show better transfer to independent work; AI expertise (from Linear explanations) will show better error source identification

Implementation: Online Learning Game

Prolific Study Design

- ~200 participants across 5 conditions
- Performance-based bonuses (accuracy + error detection)
- 25-30 minute engaging game format

Quality Assurance

- Multiple attention checks during learning phases
- Comprehension verification of game rules
- Device restrictions for consistent visualization

Measurement

- Real-time learning curve analysis
- Transfer testing without model access
- Individual difference predictors of learning success

04 Expected Results

Research Insights

When Do Explanations Help?

- Do visible models accelerate learning or create cognitive overload?
- Does this depend on model complexity (linear vs. GAM)?
- Are there individual differences in who benefits from transparency?

Reverse Engineering AI Behavior

- Can people learn to predict AI outputs without seeing explanations?
- How long does this implicit learning take?
- What patterns are learnable vs. too complex to infer?

Error Detection and Calibration

- Do people learn when to trust vs. override the AI?
- Which conditions best teach AI reliability boundaries?
- How does explanation visibility affect confidence calibration?

Contributions

Theoretical

- Understanding human learning from AI explanations
- Extension of cognitive load theory to repeated interaction contexts
- Individual difference predictors of explanation effectiveness

Practical

- Guidelines for when to provide explanations vs. predictions only
- Understanding of learning timescales for different explanation types
- Design principles for AI systems that support user learning

Methodological

- Validated approach for measuring learning from AI (not just understanding)
- Performance-based evaluation paradigm for explanation research

05 Discussion

Key Challenges

Dual Learning Effectiveness

- Do GAM explanations better support **domain learning** while linear explanations better support **AI behavior learning**?
- How many rounds are needed to see **meaningful differences** in each learning type?
- What **individual differences** predict success in domain vs. AI expertise development?

Practical Implications

- **How do findings generalize** to other AI application domains?
- **What are the trade-offs** between explanation complexity and learning?
- **When should real systems** prioritize transparency vs. simplicity?

Methodological

- **Are performance incentives** sufficient to motivate genuine learning?
- **How do we distinguish** learning from memorization in the transfer phase?

What We're Looking For Today

Resonance Check

- Does this shift from understanding to learning feel important?
- Is the multi-phase game structure compelling and realistic?

Critical Feedback

- What are the biggest threats to internal/external validity?
- Where might we be overcomplicating or oversimplifying?
- What essential elements might we be missing?

Constructive Input

- Suggestions for improving the learning game design?
- Better ways to measure internalization and transfer?
- Key literature or methodological approaches we should consider?

The Vision

Move beyond: "Do people understand this explanation?”

Move toward: "Do people learn to work better with AI over time?"

The opportunity: Understand how explanation design affects the gradual development of human-AI collaboration skills.

Today's goal: Get your insights on whether this learning-focused approach is promising and how to strengthen it further.

Questions & Discussion