Responsible AI Deployment in Safety-Critical Settings

A study by Morey, Rayo & Woods on empirically derived evaluation requirements for AI in healthcare

The Challenge: AI in Safety-Critical Environments

Growing AI Adoption

Increasing application of AI in safety-critical settings like digital medicine, despite polarizing debates among experts and practitioners

Historical Concerns

Digital technologies in healthcare have often introduced new risks, many recognized only after contributing to patient harm

Evaluation Gap

Both the impacts of AI augmentation and methods of evaluation have been highly variable, leaving unclear whether deployments will be helpful or harmful

Despite promising demonstrations of AI capabilities, recognition of AI errors, harms, and failed implementations have amplified calls for responsible deployment.

Study Design: Testing AI in Patient Monitoring

Participants

- 450 nursing students
- 12 licensed nurses

Task

Analyze 10 historical patient cases to recognize imminent patient deterioration

Conditions Tested

- No AI augmentation
- AI recommendations only
- AI explanations only
- Both recommendations and explanations

The study evaluated three augmentative AI technologies nurses used to recognize imminent

AI System Used in the Study

Patient Data Interface

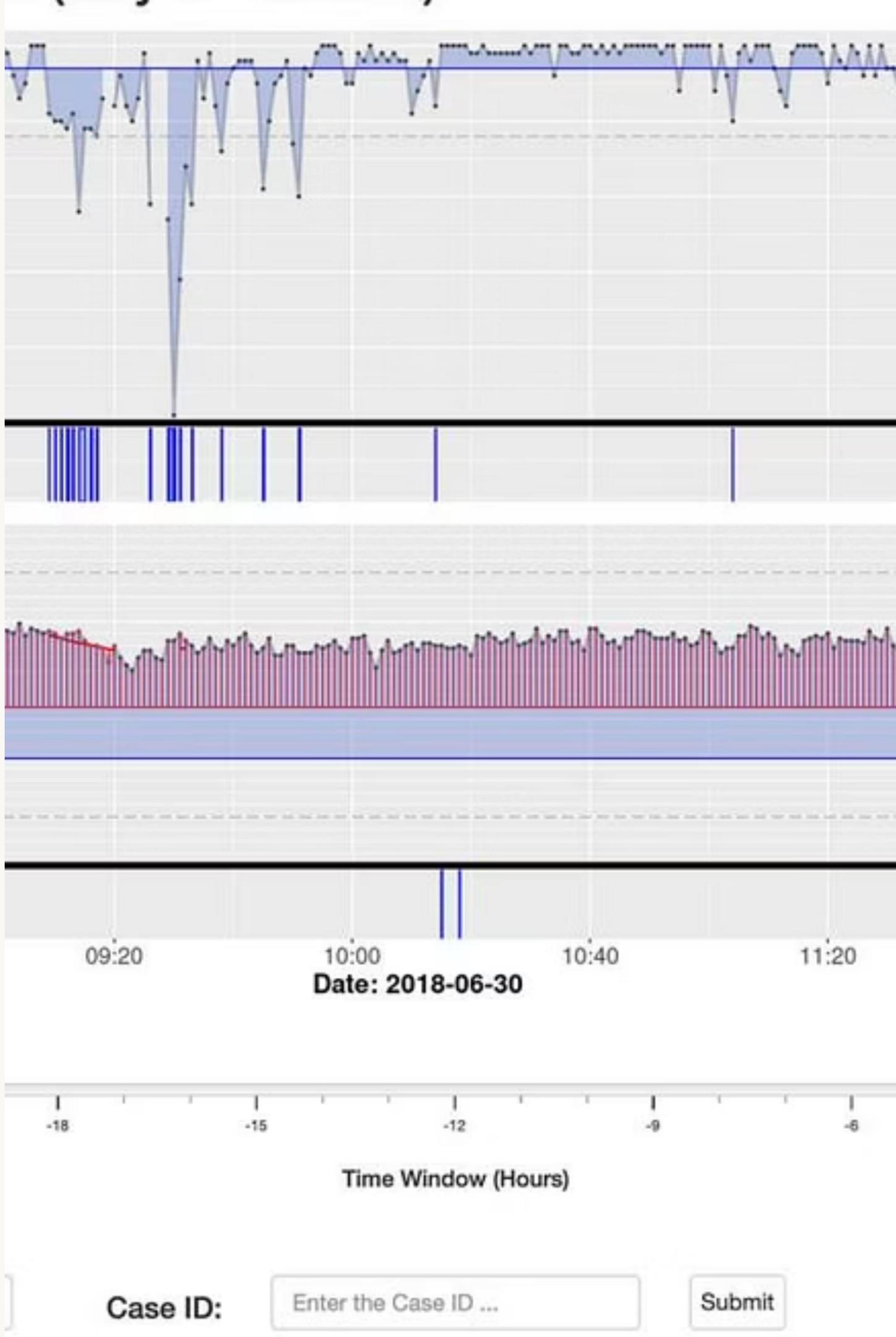
Designed to aid early recognition of patient deterioration, showing vital signs, lab results, and patient demographics

Predictive Algorithm

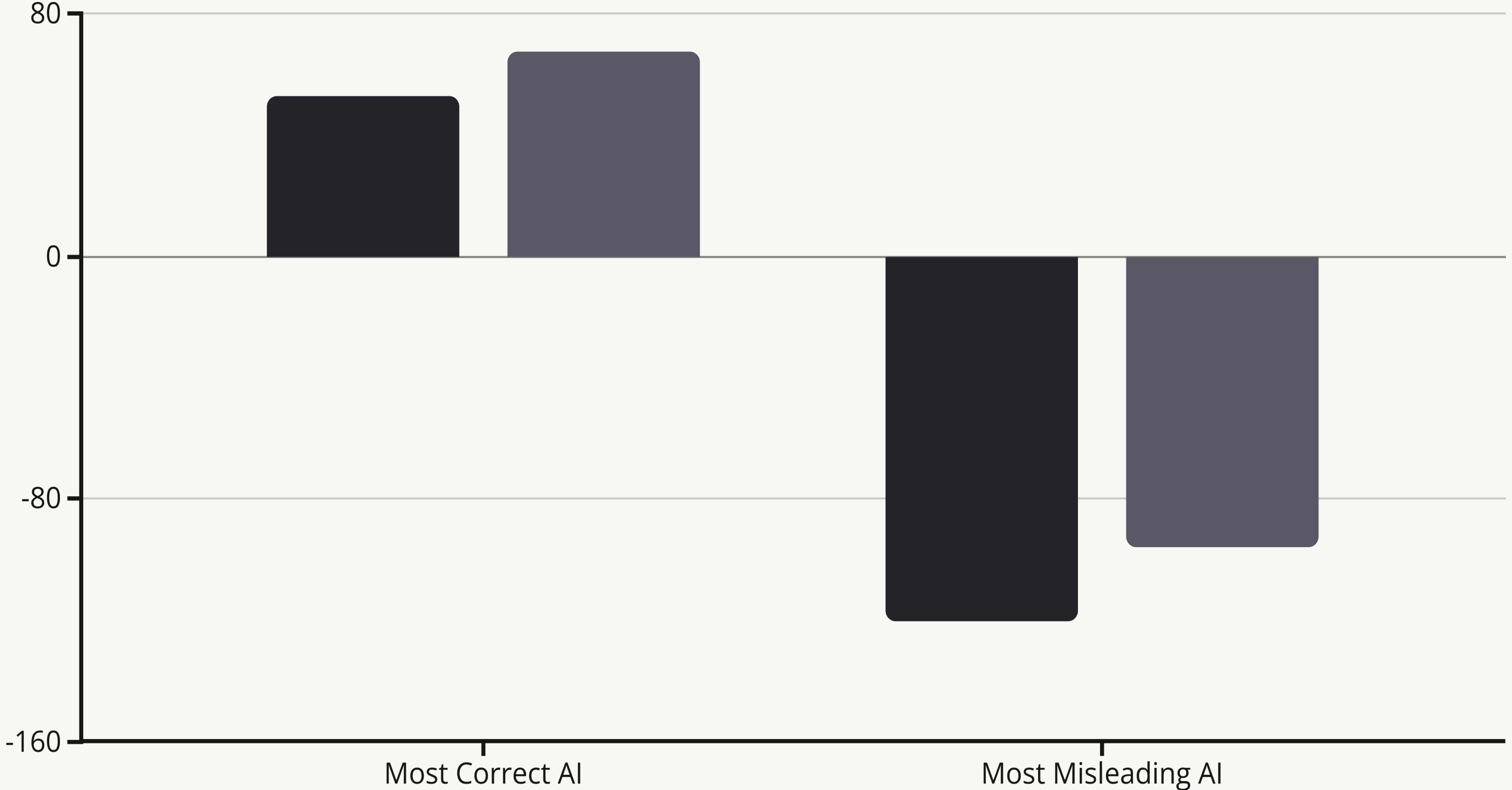
Logistic regression model with 71% accuracy, predicting probability (0-100%) of emergency events occurring in next 5 minutes

Visual Explanations

Salience-based technique highlighting which data strongly contributed to the AI recommendation



Key Findings: The Double-Edged Sword of AI



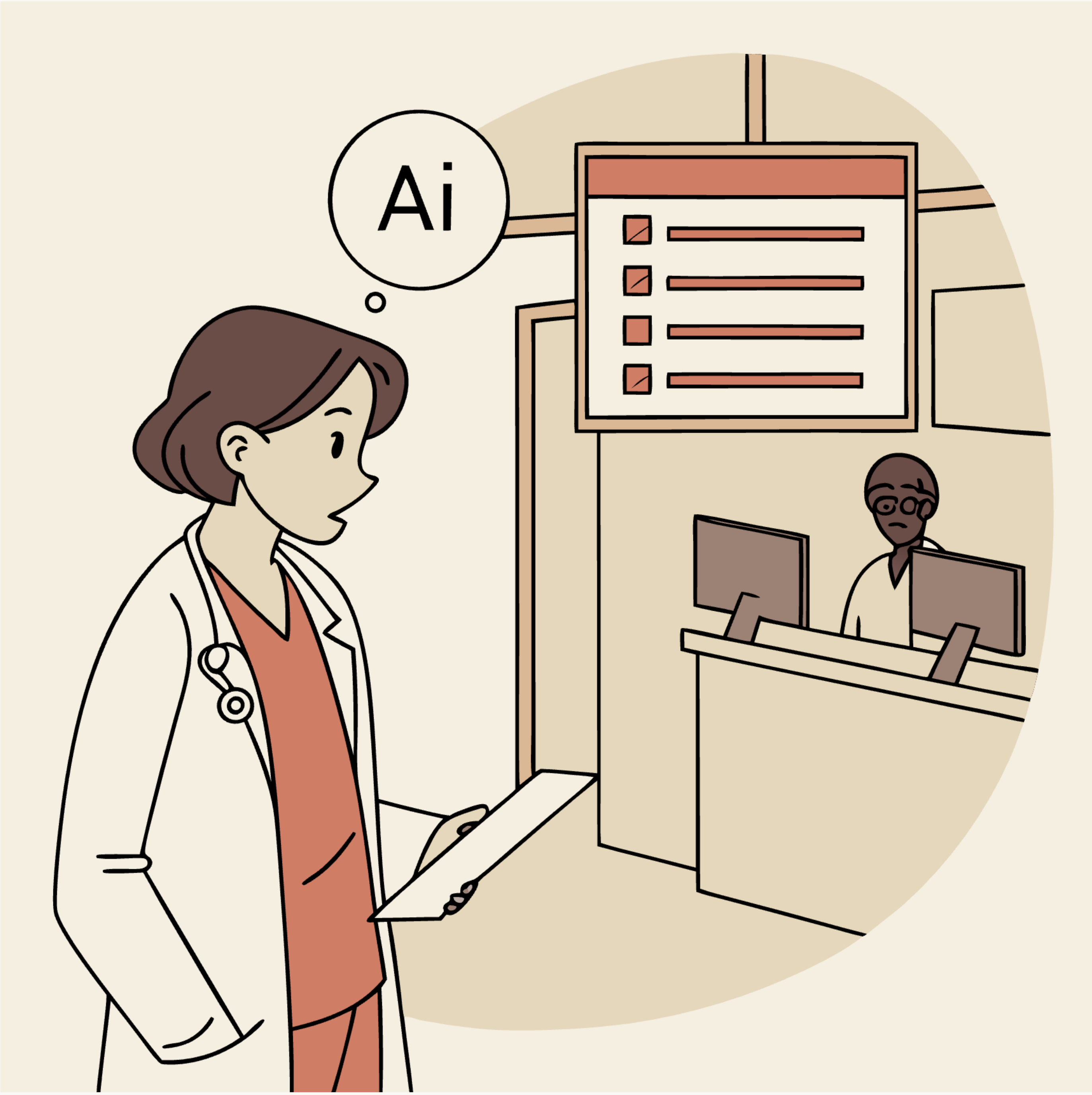
Surprising Results About AI Explanations

Explanations Did Not Mitigate Risks

The addition of AI explanations did not significantly alter results when AI recommendations were also shown. When only AI explanations were shown, the impact was smaller but followed the same pattern.

> "Our findings caution that human supervision may not fully mitigate the risks of all AI errors, even when provided with AI explanations."

This is especially surprising because the study utilized an interpretable and human-centered algorithm with explanations designed to highlight algorithm errors.



All forms of AI augmentation significantly degraded human-AI performance when the AI

Two Minimum Requirements for Responsible AI Evaluation

1

Empirically Measure People and AI Together

Evaluations must assess the joint human-AI system performance, not just AI capabilities alone. Improving AI reliability or explainability doesn't necessarily produce a safe and effective joint system.

- Few evaluations involve domain practitioners
- Separate evaluations of human and AI performance require assumptions about joint performance
- AI technologies should not be assumed to always exceed human capabilities

2

Examine a Range of AI Performance

Evaluations must include challenging cases that produce strong, mediocre, and poor AI performance to fully assess risks.

- Recognizing and recovering from AI errors is consistently challenging
- Evaluations that only consider strong AI performance are incomplete
- The frequency of errors is often underestimated in foresight

Implications for Responsible AI Deployment

Key Takeaways

- AI capabilities alone do not guarantee a safe and effective joint human-AI system
- AI augmentation can both improve and degrade performance depending on problem difficulty
- Explainable AI may be insufficient to mitigate risks of misleading recommendations
- Improving AI performance doesn't necessarily improve joint human-AI performance

Next Steps

Future research should investigate alternative human-AI architectures that might be less susceptible to propagating AI errors and examine how people interact with augmentative AI technologies over time.

Understanding potential risks before people are harmed is a fundamental responsibility of