ARL-TR-10256 ● JAN 2026

The Limits of Explanation in Human–AI Teaming: Evidence from Sports Forecasting

by Ayobami Oyewole, Steven Thurman, and Ramesh Srinivasan

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-10256 ● JAN 2026

The Limits of Explanation in Human–AI Teaming: Evidence from Sports Forecasting

Ayobami Oyewole and Ramesh Srinivasan
University of California, Irvine

Steven Thurman
DEVCOM Army Research Laboratory

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED	
		START DATE	END DATE
January 2026	Technical Report	30 April 2025	19 November 2025

4. TITLE AND SUBTITLE
The Limits of Explanation in Human–AI Teaming: Evidence from Sports Forecasting

5a. CONTRACT NUMBER	5b. GRANT NUMBER	5c. PROGRAM ELEMENT NUMBER
W911NF242001318		
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**
ARL-25-041B		

6. AUTHOR(S)
Ayobami Oyewole, Steven Thurman, and Ramesh Srinivasan

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
DEVCOM Army Research Laboratory ATTN: FCDD-RLA-FE Aberdeen Proving Ground, MD 21005	ARL-TR-10256

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.

13. SUPPLEMENTARY NOTES
ORCID: Steven Thurman, 0000-0003-3962-8447

14. ABSTRACT
This study examines whether the relevance of AI explanations improves learning and performance in hybrid human–AI binary-event forecasting. Ultimate Fighting Championship outcomes were forecast by 45 participants across 50 trials while interacting with 1 of 3 conversational agents matched in accuracy (68%): *GoodAI* (highest-SHapley Additive exPlanations [SHAP] feature), *BadAI* (lowest-SHAP feature), or *NoAI* (no rationale). Performance was incentivized wagering; learning was quantified as pre-/post-changes in subjective feature weights, compared with an optimal statistical model and with each condition's rationale distribution. Contrary to expectations, AI condition had no reliable effect on final earnings (ANOVA F (2, 38) = 0.79, p = 0.46) or on alignment to optimal weights (F (2, 38) = 2.12, p = 0.13); rationale distributions showed no selective influence on behavior. There was a small trend favoring *NoAI* in terms of mental model alignment with the optimal feature weights. In feedback-rich tasks with compact feature sets, static single-feature rationales may add little informational value. Methodologically, we contribute a relevance-controlled evaluation paradigm for human-validated explainable AI.

15. SUBJECT TERMS
artificial intelligence; human–AI teaming; explainable AI; forecasting; mental models; decision-making; Humans in Complex Systems

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES
a. REPORT	b. ABSTRACT	C. THIS PAGE		
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UU	28

19a. NAME OF RESPONSIBLE PERSON	19b. PHONE NUMBER *(Include area code)*
Steven Thurman	(520) 691-4791

STANDARD FORM 298 (REV. 5/2020)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

List of Tables

1. Introduction

Artificial intelligence (AI) agents have rapidly evolved from passive analytic tools to active teammates that share the cognitive workspace with humans in finance,[1] healthcare,[2] defense,[3] and sport analytics.[4] This transition has produced what many scholars characterize as *hybrid multi-agent decision-making*: socio-technical systems in which human forecasters and computational models contribute complementary skills, arbitrate controls, and co-construct solutions.[5] In this setup, success relies not only on the raw predictive capabilities of algorithms but also on whether human teammates can trust, understand, question, and strategically utilize machine output. The named phenomenon for such understanding is the "mental model," an internal representation of causal configuration that enables humans to anticipate how the environment and their artificial teammates will behave.[6]

1.1 Human Mental Models in Forecasting

Binary event forecasting tasks, such as predicting stock price shifts or sport outcomes, provide a tractable analytical test bed for studying mental model formation. In such tasks, forecasters observe a set of data, assign subjective weights, and formulate an overall judgment. Empirical and theoretical work shows that the accuracy of these observed data-weight maps track reasonably well onto forecasting performance.[7] Linear bootstrapping models often outperform experts precisely because they replicate an optimally weighted observed data integration process that humans approximate only with excessive training or under assistive conditions.[8] When tasks provide feedback, forecasters can update observed data weights, gradually nudging subjective models toward an objective optimum.[9] In practice, optimal mental models may not be achieved due to factors such as limited attention, noisy environments, and cognitive shortcuts that inhibit learning or divert attention to weak predictors.

1.2 AI Explanations as Learning Scaffolds

Explainable AI (XAI) research proposes that model-side explanations, feature attributions, example-based justifications, and counterfactual contrasts can catalyze human learning, reinforce trust, and mitigate inappropriate automation bias.[10,11] However, in practice, evidence for these supposed benefits is seldom seen. Several controlled experiments report gains in transparency and downstream task accuracy when explanations align with ground-truth feature importance.[12] Other studies report no improvement and sometimes even performance degradation, when explanations are ambiguous, redundant, or more problematically, direct users'

attention toward weakly predictive observed data.[13] These findings mimic the classic "seductive detail" effects found in educational psychology: providing additional information increases subjective satisfaction but distracts from core principles.[14]

The slow progress of XAI in realizing its intended benefits could be attributed to a methodological challenge. A comprehensive audit of 18,254 XAI publications showed that fewer than 1% incorporated any human evaluation in their explainability claims.[15] The vast majority of studies measure "explanation quality" by internal-model criteria, such as sparsity or fidelity, without examining whether humans actually learn from, or even understand, the provided explanations. As a result, the field lacks consensus on when explanations genuinely help users form more accurate mental models or when they merely amplify noise. Bridging this gap requires experimental paradigms that 1) hold algorithmic accuracy constant, 2) systematically manipulate explanations *relevance*, and 3) capture both *internal* learning (mental-model change) and *external* performance (forecast accuracy).

In this study, we address these gaps with an experimental paradigm in the domain of Ultimate Fighting Championship (UFC) bout prediction, a high-variance environment familiar to many participants yet unpredictable enough to exclude ceiling effects. Each fight bout is described by a vector of paired interpretable statistics for the fighters (wins, losses, age, height, striking metrics, and defense metrics), enabling transparent mapping of features and outcomes. Previous work shows that sport-analytics tasks birth rich strategy formation and are sensitive to framing manipulations, making them ideal for mental-model studies.[16]

To isolate the impact of explanation relevance, we designed three conversational AI agents, all calibrated to equal classification accuracy (68% on held-out historical UFC data) but differing in their explanatory strategy: ***GoodAI*** highlights the most informative (highest-SHapley Additive exPlanations [SHAP]) feature for the bout and verbally links it to the predicted winner; ***BadAI*** highlights the least informative (lowest-SHAP) feature as the explanation for the predicted winner; and ***NoAI*** provides no feature-based explanation as a control for task learning in the absence of AI support. This design follows calls for *relevance-controlled* XAI experiments in which explanatory information is orthogonal to predictive accuracy.[17] We compare GoodAI with BadAI while benchmarking against NoAI, to test whether explanatory alignment alone can drive mental-model evolution and downstream wagering behavior. Grounded in cognitive theory and XAI practice, we pose two research questions (RQs):

RQ1: How does the relevance of an AI explanation shape the refinement of human mental models during iterative binary event (UFC bout) forecasting?

RQ2: How does explanation relevance impact forecasting accuracy and task-induced learning of optimal feature weights?

In relation to these questions, we formulated the following two hypotheses (H) to guide the design and analysis of this research study.

H1: Participants' mental models of feature importance will be influenced by the frequency and distribution of AI explanations across trials such that participants in the GoodAI condition will develop more statistically optimal mental models (aligned to the highest SHAP-valued features), while participants in the BadAI condition will develop suboptimal mental models (aligned to the least predictive features with the lowest SHAP values). Participants in the NoAI condition will serve as a control for unbiased task-based learning of feature weights in this task without AI influence.

H2: Mental model alignment will impact performance by directing attention toward relevantly (or unrelevantly) weighted observed data such that GoodAI will increase trial-level prediction accuracy and total wager earnings compared to NoAI, whereas BadAI will reduce both metrics by misallocating attention and cognitive resources to weak predictors.

1.3 Contributions to XAI and Human–AI Teaming

This study contributes to the literature in four ways. First, it addresses the empirical vacuum identified by Suh et al.[15] by providing a fully transparent human evaluation of AI explanations in a naturalistic forecasting task. Second, it decouples predictive accuracy from explanatory quality, allowing clean causal inference about the direction of learning. Third, it leverages self-reported feature importance ratings pre- and postexperiment to quantify changes in participants' mental models, extending recent methodological guidelines for XAI evaluation.[18] Finally, it offers actionable design principles: conversational agents that highlight relevantly *aligned* features to scaffold learning and domain understanding, in contrast to other approaches that may corrode expertise over time.

2. Related Works

Research on human decision-making has long treated forecasting as a policy learning problem in which humans map observed data to outcomes using internally maintained weighting schemes. An underlying implication is that these observed data-weight "policies" can be obtained, compared to the environmental structure, and evaluated for consistency and bias.[7] Meta-analytic evidence shows that simple linear bootstrapping models often outperform human forecasters because they

apply stable, optimally weighted integration, spotlighting where human subjective policies typically fall short.[8] The mental-model mapping tool "M-Tool," a recent assessment technique, provides scalable procedures to uncover and track people's causal beliefs, enabling a longitudinal study of the evolution assessment of mental models with exposure and practice.[6] Notably, timely and informative feedback drives learning environments toward the positively incremental end of the spectrum, where forecasters can adjust observed data weights and progressively align their policies with the objective structure.[7] Together, these insights informed our decision to use pre/post mental-model similarity as primary outcome measures.

Beyond offering abundant and easily interpretable features, sports settings present repeated high-variance events that evoke rich strategic adaptations. Evidence from experimental and real-market studies shows that small changes in presentation or compensation can induce large shifts in wagering strategies, revealing that forecasters' internal models are sensitive to framing and feedback.[16] In addition, the broader literature on sports analytics documents how data-driven methods are used to model tactical choices and performance trade-offs, underscoring the domain's suitability to examine how explanations steer attention and observed data weighting.[4] These characteristics make prediction of sport outcomes an ecologically valid yet tractable analytical test bed for studying explanation-driven learning.

Today's decision systems can be viewed as *hybrid-intelligence* teams, where humans and algorithmic models contribute complementary competencies, arbitrate control, and co-construct solutions.[5] Empirical studies that span finance,[1] healthcare,[2] defense,[3] and sport analytics[4] show that collaborative effectiveness depends on joint process alignment (e.g., when to consult the model, and how to contest it) and role clarity (who integrates which observed data). Consequently, studies operationalize "collaborative-ness" not only through outcome metrics (accuracy, risk reduction, and payoff), but also through process measures such as advice-taking rates, turn-taking symmetry, and calibration between human confidence and model confidence. In this study, we set *explanation-relevance* as a design variable that can scaffold or impede these collaborative dynamics by changing how humans weight the evidence.

The XAI literature proposes explanations, feature attribution, example-based justifications, and/or counterfactuals to ensure that model reasoning is legible and contestable. Recent studies have also called for a rigorous science of AI interpretability with a foundation in human-centered goals,[10,11] and design research offers practical heuristics for developing user-facing rationales that answer task-relevant questions.[18] However, empirical evidence for the proposed benefit of XAI is mixed: aligned explanations are said to improve downstream judgments in some

tasks,[12] yet another study reported that visual or textual rationales can also misdirect attention toward weakly predictive observed data or promote over-reliance.[13] Moreover, explanation policies that ignore the underlying distributional impacts of the training dataset can inadvertently produce dissimilar advice-taking across users.[17] These findings suggest that the explanation *relevance*-connection to strong features of the current decision is a more actionable lever than format alone.

Despite the progress reported on XAI by its propagators, systematic evaluation of explanations with human participants is very rare: a recent audit of more than 18,000 XAI papers found that fewer than 1% of them include human validation.[15] Even when humans are studied, evaluations often prioritize *attitudes* (e.g., trust) over *learning* and combine explanation quality with model competence, leaving uncertain whether explanations reshape internal observed data weighting in ways that translate into better performance. Our study aims to address this gap by maintaining the same predictive accuracy of the model across three conversational agents while manipulating the explanation *relevance*. We then jointly quantify 1) mental-model evolution through pre/post similarity to ground-truth feature weights (derived from a statistical model fit to a larger dataset), and 2) forecast success via accuracy and wagering outcomes in a high-variance UFC domain. This design isolates the causal effect of explanation alignment on both internal representations and collaborative performance.

3. Methodology

Our methodological approach comprises three stages aligned with the RQs. First, we establish a hybrid decision-making paradigm in which human forecasters interact with one of three conversational AI agents matched in predictive accuracy but differing in explanation relevance, enabling a straightforward between-subjects test of explanatory alignment. Second, we curate a trial set that remains representative of the broader UFC corpus while allowing precise control over the information each agent highlights, isolating effects of explanation content from bout idiosyncrasies. Third, we deploy an online behavioral experiment that mirrors real-world sports wagering to capture learning dynamics and performance outcomes. The subsections that follow detail the task, trial and AI design, data collection, and measurement strategy.

3.1 Experimental Task

We implemented a hybrid decision-making paradigm in which human participants make trial-by-trial forecasts while observing a conversational AI agent's prediction about UFC bout outcomes. Our design follows best practices from incentivized

forecasting tournaments, which use repeated judgments, immediate feedback, and performance-contingent bonuses to sustain engagement and elicit effortful, policy-level learning.[19] On each trial, participants place a wager in discrete units from 0 to 4 that reflects their confidence in the AI prediction, and they receive immediate feedback about the realized outcome. The experiment proceeds in five stages, namely, pretask, onboarding, task familiarization (10 practice trials), main experiment (50 trials), and posttask, to scaffold familiarity with the interface, stabilize task understanding, and permit measurement of both learning dynamics and performance under repeated decisions (see Figure 1 for a schematic flow).

In the pretask stage, participants observe the full set of fighter features (with a description of what each feature represents), and use a slider bounded from 1 to 100 to indicate the perceived importance of each feature for predicting fight outcomes. Progression is gated so that each slider must be adjusted at least once, the default value is 1, ensuring complete coverage of subjective priors. The onboarding stage then provides a guided walk-through of the interface, including the location and format of feature presentations, the wallet display, the panel in which the AI explanation appears, the wagering control, and the outcome display. A practice block follows, during which participants complete a set of 10 bouts with the identical interaction flow used later. Practice data are excluded from analysis.

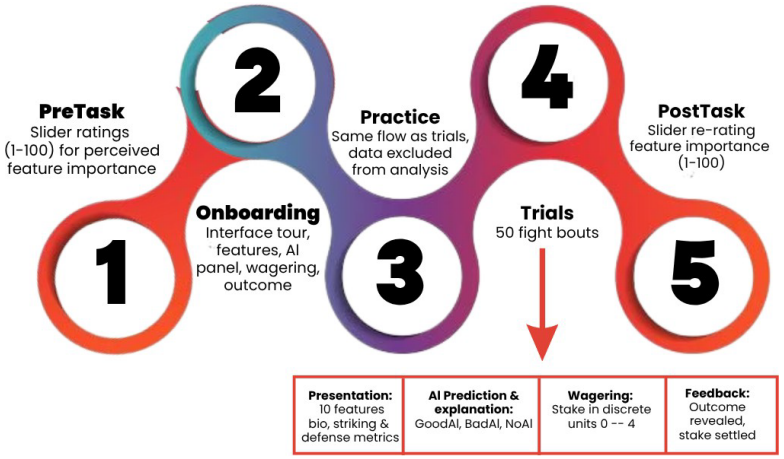

Figure 1. Diagrammatic flow of the experimental setup. Participants proceed from pretask to onboarding to practice, then complete 50 trials comprising presentation, AI prediction and explanation, wagering with a 0–4 stake, and feedback, and finally, complete posttask re-elicitation.

During the main experiment, each bout presents 10 interpretable features balanced across fighter biographical information, striking metrics, and defensive metrics. Participants begin with a virtual bankroll of $100. They observe the AI's categorical prediction (Fighter A or Fighter B), and, depending on the condition, observe an accompanying explanation such as, *"Fighter A is predicted to win due to more*

career wins than Fighter B." They then place a stake between 0 and 4 units and receive outcome feedback that updates the bankroll, which can grow or shrink over time. To approximate real-world incentives, participants receive a performance-based bonus in addition to a fixed participation fee, thereby aligning payoffs with task success and sustaining attention across the 50-bout sequence.

Explanation relevance is manipulated between participants with three conditions. GoodAI emphasizes the feature with the highest predictive value for the current bout, BadAI emphasizes a weakly predictive or irrelevant feature, and NoAI provides no feature-based rationale. This structure allows explanation content to modulate attention and wagering strategy while holding predictive accuracy fixed.

3.2 Trial Selection and AI Design

We began with the publicly available UFC Complete Dataset spanning events from 1996 to 2024, which carries a reported usability score of 10.0 on Kaggle and provides bout-level statistics and fighter attributes suitable for supervised learning.[*] The raw corpus contained 7439 bouts; after removing duplicates and records with impossible values, for example, zero height or missing age, 5422 bouts remained and were exported as a cleaned comma-separated file for analysis. From the full schema, we down-selected 10 fighter attributes to balance interpretability with coverage of performance dimensions. The selection included a balance of four types of variables: historical performance, offensive capability, defensive resilience, and individual characteristics. Concretely, we retained three offensive statistics, such as ***significant strikes landed per minute***, ***takedown accuracy***, and ***striking accuracy***, three defensive statistics, including ***striking defense***, ***takedown defense***, and ***significant strikes avoided per minute***, two demographic attributes, ***age*** and ***height***, and two historical indicators, ***total wins*** and ***total losses***. We deliberately excluded weight because of ceiling effects within divisions that compress variability, omitted reach in favor of height to avoid collinearity while keeping an interpretable proxy for body stature, and dropped categorical variables, such as stance, to maintain a compact, encoding-free feature set.

Model training followed a prespecified pipeline. We partitioned the cleaned data into an 80 to 20 training and test split at the bout level, yielding 4337 training bouts and 1085 test bouts. We then engineered fight-specific difference features so the model consumed a single vector per bout, with signs chosen so that lower is better for age and losses. To reduce feature scale bias, we applied a two-step normalization fitted on the training partition, standardization to zero mean and unit

[*]Basharov M. UFC complete dataset, all events 1996–2024. Kaggle: https://www.kaggle.com/datasets/maksbasher/ufc-complete-dataset-all-events-1996-2024

variance followed by a minimum-to-maximum rescaling to the zero-to-one range and propagated those parameters to the test set. To address class imbalance we applied SMOTE on the scaled training set before model fitting.[20] We estimated a logistic regression with LIBLINEAR solver and a grid over $C \in \{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100, 1000\}$ crossed with ℓ_1 and ℓ_2 penalties using five-fold cross validation. The held-out test accuracy was 0.68, which we intentionally retained to avoid a ceiling that would overshadow human wagering behavior.

AI design and trial curation were coupled to preserve predictive competence while manipulating explanation relevance. From the test partition we drew 100,000 random candidate sets of 50 bouts and retained those whose simulated accuracy fell within ± 0.01 of the model's overall test accuracy, yielding 12,359 candidate sets. For each candidate set we produced GoodAI and BadAI rationales using a small library of natural language templates and a SHAP explainer fit on the trained model with a background of 100 randomly sampled training instances.[21] Representative templates are summarized in Table 1. For a given bout we computed SHAP values per feature for the predicted winner, identified the leading features, then selected the highest absolute attribution for GoodAI and the lowest absolute attribution within the leading set for BadAI, treating the most important feature for the nontarget class as the most negative attribution in absolute value. To choose a single trial set for the study, we compared each candidate's regression coefficients estimated on its ground truth labels and on its predicted labels, $\boldsymbol{\beta}_{\text{gt}}$ and $\boldsymbol{\beta}_{\text{pred}}$, to the full model coefficients $\boldsymbol{\beta}_{\star}$, computed the distribution vector $\mathbf{s}$ of GoodAI rationale features, and minimized the composite score in Equation 1,

$$\text{Score} = \text{MSE}(\boldsymbol{\beta}_{\star}, \boldsymbol{\beta}_{\text{gt}}) + \text{MSE}(\boldsymbol{\beta}_{\star}, \boldsymbol{\beta}_{\text{pred}}) + \text{MSE}(\boldsymbol{s}, \boldsymbol{\beta}_{\text{pred}}), \text{ where MSE}(\mathbf{a}, \mathbf{b}) = \frac{1}{d}\sum_{j=1}^{d}(a_j - b_j)^2 , \quad (1)$$

where d is the number of features. The selected file was trial set 3761.csv with a score of 0.0037, which provided close alignment to the global model while preserving diversity in rationale features.

Table 1. Textual justification templates used to instantiate GoodAI and BadAI explanations.

Lower-better rationale	High-better rationale
Features: age, total losses {fighter} is favored because it has a lower {feature} compared to its opponent.	Features: striking accuracy, striking defense {fighter} is favored because it has a higher {feature} indicating superior performance.
{fighter} is more likely to win due to its lower {feature}, indicating superior defense or youth.	{fighter} is more likely to win thanks to its higher {feature} compared to its opponent.
A lower {feature} gives {fighter} a competitive edge in the fight.	A higher {feature} gives {fighter} a competitive advantage in the fight.

NOTE: The left column applies to features where lower values are favorable, for example, age and total losses, and the right column applies to features where higher values are favorable, for example, striking accuracy and striking defense. Placeholders {fighter} and {feature} are replaced with the predicted winner and the selected rationale feature.

3.3 Data Collection

We implemented the experimental task as a UFC web-based game built with HTML, CSS, and JavaScript. The application was hosted on Amazon Web Services to enable shared access, and all event level data were written to a Firestore database in real time, including time stamps, trial identifiers, feature presentations, AI condition, wagers, outcomes, and bankroll updates. Participants were recruited through Prolific and redirected to the hosted task and returned to Prolific for compensation processing. Each participant received a fixed payment of $15 and a performance-based bonus of $0.10 for every dollar earned above $168, which corresponds to the average return a participant could obtain by deferring to the model prediction throughout. The study lasted approximately 60 min per participant. We enrolled 45 participants and randomly assigned 15 to each AI condition, yielding balanced cells for analysis. Figure 2 summarizes the principal participant facing screens, panel (a) shows the slider interface used for both the pre- and posttask feature importance elicitation, and panel (b) shows the trial interface with wallet, bout information, AI prediction and explanation, and the wager control.

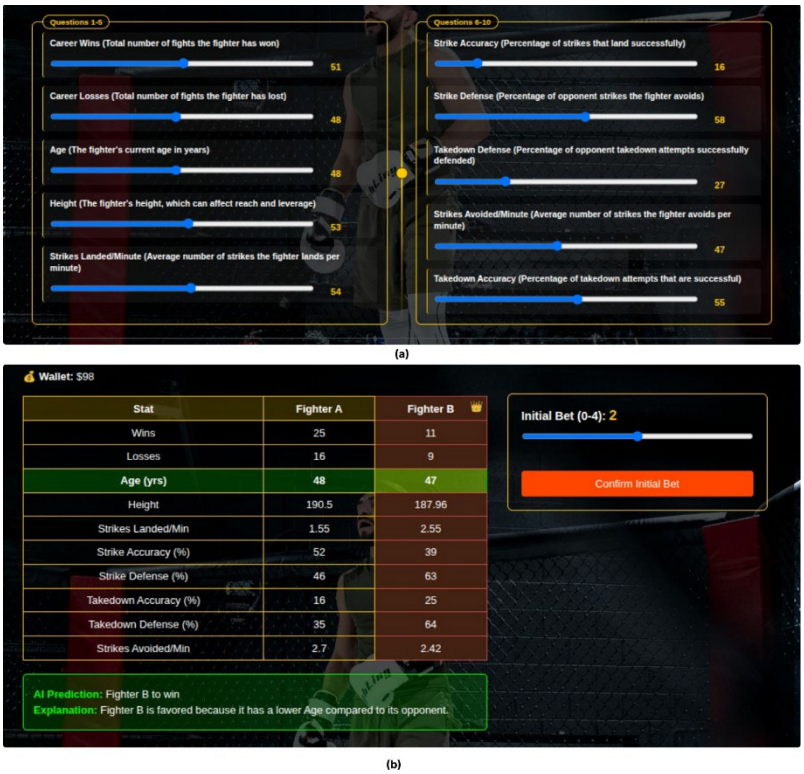

Figure 2. Screens from the web-based task. Panel (a) slider interface used for both the pre- and posttask feature importance elicitation. Panel (b) trial interface showing wallet, bout information, AI prediction and explanation, and the wager control.

The protocol received human subjects' approval from the University of California, Irvine Institutional Review Board (IRB Reference Number 6933), and from the U.S. Army Combat Capabilities Development Command (DEVCOM) Army Research Laboratory (ARL 25-041B). The first screen of the application presented an informed-consent statement describing study procedures, risks, benefits, compensation, and data handling, after which participants could decline and exit to Prolific or provide consent and proceed. Participants were required to affirm that they had read and understood the information before continuing, and the onboarding sequence included a brief instruction check to confirm attentiveness to the task rules and interface. The consent flow and quality control checks preceded all task interactions to ensure ethical participation and reliable data.

3.4 Measurements

Participant interactions during the main experiment, and during the pre- and posttask surveys, provided behavioral data as the basis for this analysis. Overall task performance was measured by final wallet value, reflecting the confidence and success of wagers made across trials. Optimizing performance would require participants to calibrate their wagers such that high-value bets (3–4) would be made

on AI-correct trials (0.68 probability) and low-value bets (0–1) made on AI-incorrect trials (0.32 probability). The final wallet value thus provides a cumulative measure of overall task performance.

The pre- and posttask surveys elicited subjective ratings of feature importance across the 10-feature space using sliders that ranged from 1 to 100 units. These ratings served as an estimate of the internal mental model of the participant regarding the relative importance attributed to each feature. The mental-model feature weights derived from slider data were normalized to sum to one by dividing each by the sum of all 10 slider values. This data transformation preserved the relative feature weights across the set but allowed for standardized comparisons between pre- and post-assessments, and across individuals. The mental model, therefore, represents a normalized vector of relative weights associated with each of the 10 fighter features ranging between 0 and 1.

4. Analysis and Results

To assess the influence of AI condition on task performance, we performed a one-way analysis of variance (ANOVA) on final wallet value, revealing that there were no statistically significant differences among the three conditions (F (38, 2) = 0.79, p = 0.46). As seen in Figure 3d, the time course and trajectory of wallet value across the 50 trials was quite variable across individuals, but very similar between the three groups. Mean final wallet value was 142.4 (std = 14.2), 148.4 (std = 18.3), and 140.5 (std = 19.3), for the GoodAI, BadAI, and NoAI conditions, respectively.

To assess the specific influence of AI explanations on mental models, the normalized weight vectors were compared to the distribution of AI rationales in the GoodAI and BadAI conditions, representing the relative frequency that each feature category was used as an explanation across all trials of the experiment (Figure 3a). GoodAI rationales were derived from the feature with the highest SHAP value associated with the trained model—hence, the distribution of rationales resembled very closely the statistically optimal feature weights (comparing Figure 4a and Figure 3a). In this way, the GoodAI explanations should, over time, influence participants to focus on the most informative features. By contrast, BadAI rationales were derived from the feature with lowest SHAP value and differed substantially from the GoodAI distribution and optimal model weights. The BadAI explanations should, by contrast, influence participants to focus on less informative features.

We sought to capture the influence of AI explanations by computing the root mean square error (RMSE) of differences between mental-model weights and the relative frequency of rationales in each AI condition. If there was an influence of AI

explanations on mental models, we would expect lower RMSE values with the GoodAI frequency distribution for GoodAI group and higher RMSE values for the BadAI group. Conversely, we would expect lower RMSE values with the BadAI frequency distribution for the BadAI group and higher RMSE values for the GoodAI group. The NoAI group should have no particular bias because they did not receive AI explanations during the task. A schematic of this pattern of predicted results is shown in Figure 3b.

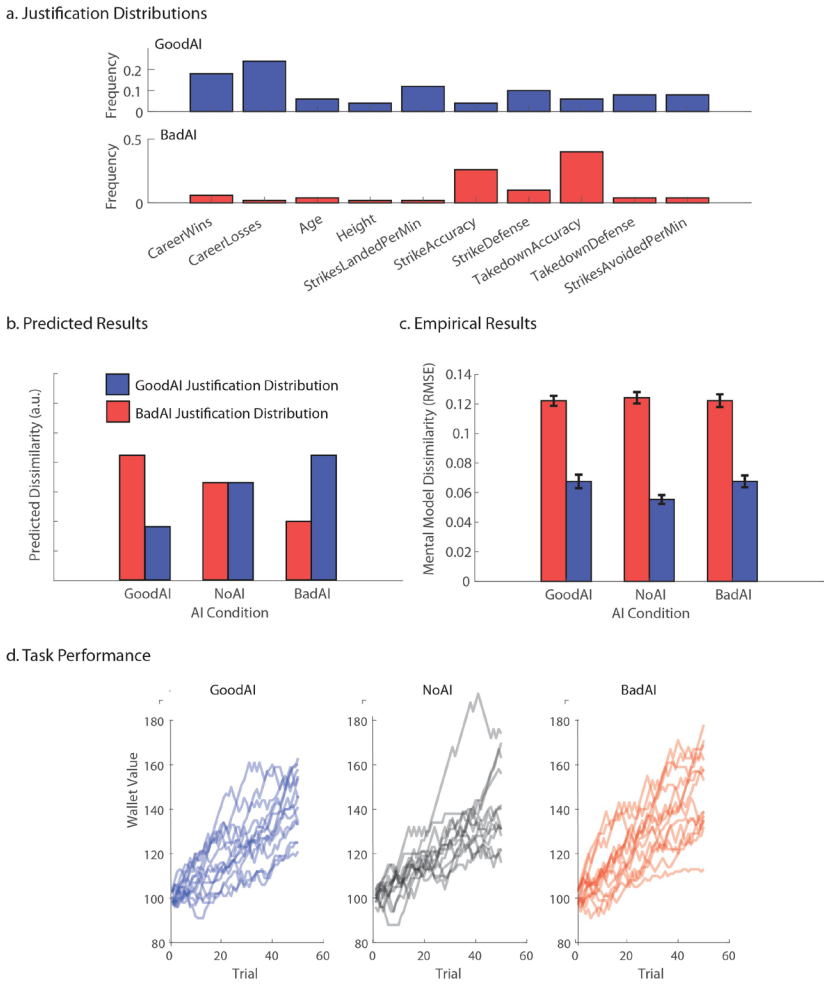

Figure 3. Normalized distributions showing the frequency of AI explanations across feature space (a) for the GoodAI condition (top) and BadAI condition (bottom). Predicted pattern of results on RMSE if AI explanations influenced participants' mental models (b). Study results showing mental-model dissimilarity (RMSE) between self-reported feature weights and the explanation distributions of GoodAI (blue) and BadAI (red) (c). Task performance over time, showing total wallet value over time (50 trials) for each condition (d).

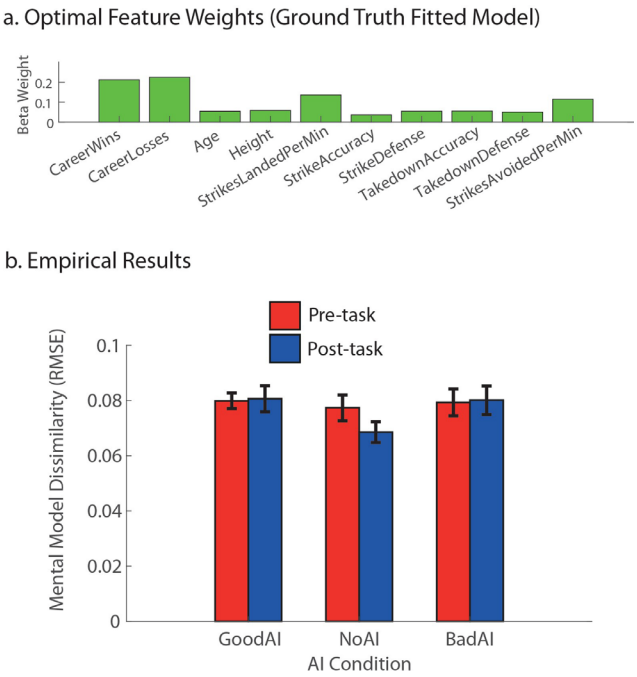

Figure 4. Normalized beta weights associated with the multiple logistic regression model fitted to a large corpus of historical UFC data (4337 fights), representing the statistically optimal feature weights for the task (a). Study results showing the mental-model dissimilarity (RMSE) for each AI condition compared to the optimal beta weights, showing the similarity of the pre- and posttask mental-model estimates (b).

Contrary to this prediction, we found no evidence that mental models were influenced by AI explanations. One-way ANOVAs revealed no significant difference among the three groups with respect to the GoodAI rationale distribution (F (38, 2) = 3.07, p = 0.058), and the BadAI rationale distribution (F (38, 2) = 0.09, p = 0.91). There was a nearly significant effect for the GoodAI distribution, highlighting that the NoAI group actually showed better alignment with the GoodAI distribution (mean RMSE= 0.055, std = 0.011) than the GoodAI (mean RMSE = 0.068, std = 0.017) and BadAI (mean RMSE = 0.067, std = 0.015) groups. There was no evidence to suggest that either AI condition was specifically influenced by the frequency of rationales and the pattern of AI explanations.

To assess the optimality of participants' mental models, the normalized weight vectors were compared to normalized beta weights derived from a statistical model (multilogistic regression) fit to the 80% training corpus of data (4337 fights). The statistical model is the same used for bout level prediction, and the beta weights are obtained by taking the absolute value of the model coefficients.

The beta weights of the fitted model were then normalized by dividing by the sum of all beta weights across the 10 features, so the optimal feature vector summed to one (Figure 4a), analogous to the normalization procedure for participants' mental models. With participant mental models and optimal feature weights normalized to

the same scale, we computed the RMSE of feature weight differences to quantify mental-model dissimilarity. Lower RMSE values indicate a closer match and more similar representation of feature weights between participants and the statistically optimal model fit to ground truth historical data.

Similar to above, we found no evidence that AI justifications helped to support (or inhibit) alignment of posttask mental models to the optimal feature weights. A one-way ANOVA revealed no significant differences among AI conditions (F (38, 2) = 2.12, p = 0.13), as shown in Figure 4b. There was a trend, again, for the NoAI group to have better-aligned mental models than either of the groups with AI rationale support. In fact, combining the GoodAI and BadAI groups into a super-group (AI Present) and comparing to NoAI directly, we found evidence for significantly lower RMSE values and more optimal mental-model alignment in the NoAI group (t (39) = 2.09, p = 0.044). Taken together, the results of this study show no evidence that AI rationales influenced behavior or performance—if anything, AI rationales (whether good or bad) may have inhibited task learning and resulted in less-aligned mental models.

5. Discussion

We found no evidence that explanation relevance contributes to the improvement of either internal learning of task-relevant feature weights or decision-making performance. With respect to RQ1, we observed that posttask mental models did not differ between GoodAI, BadAI, and NoAI compared to the distribution of rationale features or against the optimal feature weights derived from the statistical model, and the only observable trend was a closer alignment for the NoAI group compared to the AI-present groups. With respect to RQ2, final wallet value did not differ by condition and the trajectories across 50 trials were similar across conditions. Consequently, these results do not support H1 and H2 as stated and suggest that providing rationales, whether aligned or misaligned, did not measurably reshape feature weighting or improve task returns under the present design.

Several factors may plausibly explain these results. First, participants viewed the same 10 features that underline model training, which may have prompted them to form their own weighting schemes that discarded the agent's rationale, particularly in a sparse feature space. Second, the environment provides immediate outcome feedback on every trial, which tends to dominate weaker instructional signals, leading participants to rely on their own updating process rather than on textual justifications. Third, the wagering interface used a coarse 0 to 4 stake scale, which may have limited observable variance in confidence and attenuated sensitivity to

trial-wise changes in rationale. Fourth, the model accuracy was moderate, which reduces advice-taking and possibly reduces the perceived value of explanations. Finally, as an online study, the variability in attention and engagement levels could have diluted treatment effects, including limited wager variability for some participants that prevented the reliable inference of behavioral mental models from staking patterns.

Relative to the literature, these findings align with reports that the benefits of explanations are context dependent and often fragile, and contrast with studies that show gains when rationales are tightly integrated with task demands and user needs.[10–13] The study contributes a methodic human-validated test of explanation relevance in a feedback-rich forecasting task, responding to calls for empirical evaluation of XAI with users.[15] Furthermore, the slight trend toward better performance in the NoAI condition raises intriguing questions about the role of cognitive load. Explanations, even seemingly helpful ones, introduce additional information that must be processed, potentially diverting cognitive resources away from the primary task of integrating observed data. This connects to theories of cognitive load management, which emphasize the importance of minimizing extraneous cognitive demands to optimize learning and performance. It is possible that the effort required to process the AI's explanations, even the "GoodAI" explanations, outweighed any potential benefits in our experimental setting.

The results also have implications for the burgeoning field of algorithmic aversion —the tendency for people to distrust algorithms even when they outperform human judgment. Our study suggests that poorly designed explanations might exacerbate algorithmic aversion. If explanations are perceived as unhelpful or misleading, users may become even more skeptical of the AI's predictions and less likely to incorporate them into their decision-making process. This underscores the need for explanations that are not only accurate but also perceived as accurate and valuable by the user.

Looking beyond XAI, this work also connects to research in behavioral economics and decision-making under uncertainty. The UFC prediction task shares similarities with real-world forecasting scenarios, such as financial markets or geopolitical risk assessment, where individuals must make judgments based on incomplete and noisy information. The observed reliance on subjective weighting schemes, even in the face of AI assistance, highlights the persistent influence of cognitive biases and heuristics in human decision-making. The fact that immediate feedback dominated the impact of explanations also aligns with reinforcement-learning principles, where timely rewards and punishments are more effective than abstract explanations in shaping behavior.

There are notable limitations to this study, including 1) showing the full feature set may have encouraged participants to ignore the agent, 2) the 50-trial count and sample size may limit statistical power, 3) the online setting may have included inattentive respondents despite screening, and 4) reliance on subjective slider rankings is imperfect, particularly when wager data lack sufficient variation to infer behavioral weights. Ultimately, this work suggests that effective XAI design requires moving beyond simply providing explanations to carefully considering how those explanations integrate with the user's existing knowledge and cognitive processes, challenging the assumption that more explanation is always better.

6. Conclusion

This study evaluated whether the relevance of AI explanations improves learning and performance in a hybrid human–AI forecasting task. We implemented a controlled paradigm in which participants predicted UFC outcomes while interacting with one of three conversational agents that were matched in predictive accuracy and differed only in explanation relevance. We measured performance with an incentive-driven wagering scheme and assessed the change of mental model with pre- and post-subjective feature weights compared to a statistical model. Across conditions, we observed no reliable gains in wallet outcomes or in alignment of posttask mental models with either optimally informative features or the distribution of rationale features, and a modest trend favored the NoAI group. These findings suggest that in feedback-rich settings where users see the same sparse feature space as the model, static single-feature rationales may add little informational value and can even diffuse attention, reinforcing the need to evaluate explainability with human outcomes rather than model's internal mechanisms. Methodologically, the work contributes a relevance-controlled design that holds predictive accuracy constant, a principled procedure to curate trials and rationales, and metrics for evaluating internal mental models that other researchers can adopt.

This study can be extended in the future by increasing the leverage of the value-add mechanism by restructuring information asymmetry, learning time, and experimental control. First, a hidden profile design, in which the model has access to feature space that participants do not, or it summarizes patterns from data that participants cannot feasibly integrate, would make explanation content consequential and allow a direct test of whether rationales transmit privileged information that improves human policies. Second, a longer practice phase followed by hundreds of trials would provide more stable learning curves and substantially greater statistical power, which is critical to detect condition effects on both behavior and mental-model refinement. Third, an in-person laboratory implementation would permit closer monitoring of attention and strategy use, for

example, through observation and fine-grained measurement (i.e., of attention with eye tracking), thereby reducing noise from inattentive respondents that can arise in online studies. Together, these extensions would enable a stronger test of when and how explanations support human learning and decision quality in collaborative forecasting.

7. References

1. Sachan S, Almaghrabi F, Yang J-B, Xu D-L. Human–AI collaboration to mitigate decision noise in financial underwriting: a study on FinTech innovation in a lending firm. IRFA. 2024;93:103149. https://doi.org/10.1016/j.irfa.2024.103149

2. Knop M, Weber S, Mueller M, Niehaves B. Human factors and technological characteristics influencing the interaction of medical professionals with artificial intelligence–enabled clinical decision support systems: literature review. JMIR Hum Factors. 2022;9(1):e28639. https://doi.org/10.2196/28639

3. Wong JP et al. One team, one fight. Volume I, insights on human–machine integration for the U.S. Army. RAND Corporation; 2025 June 2. https://www.rand.org/pubs/research_reports/RRA2764-1.html

4. Zhou D et al. Artificial intelligence in sport: a narrative review of applications, challenges and future trends. J Sports Sci. 2025;1–16. https://doi.org/10.1080/02640414.2025.2518694

5. Dellermann D, Ebel P, Söllner M, Leimeister JM. Hybrid intelligence. BISE. 2019;61(5):637–643. https://doi.org/10.1007/s12599-019-00595-2

6. van den Broek KL, Luomba J, van den Broek J, Fischer H. Evaluating the application of the mental model mapping tool (M-Tool). Front Psychol. 2021;12:761882. https://doi.org/10.3389/fpsyg.2021.761882

7. Hogarth RM, Karelaia N. Heuristic and linear models of judgment: matching rules and environments. Psychol Rev. 2007;114(3):733–758. https://psycnet.apa.org/doi/10.1037/0033-295X.114.3.733

8. Kaufmann E, Wittmann WW. The success of linear boot-strapping models: decision domain-, expertise-, and criterion-specific meta-analysis. PLoS ONE. 2016;11(6):e0157914. https://doi.org/10.1371/journal.pone.0157914

9. Hogarth RM, Lejarraga T, Soyer E. The two settings of kind and wicked learning environments. Curr Dir Psychol Sci. 2015;24(5):379–385. https://doi.org/10.1177/0963721415591878

10. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv preprint; 2017. Ver 3 [accessed 2025 July 23]. arXiv:1702.08608. https://doi.org/10.48550/arXiv.1702.08608

11. Miller T. Explanation in artificial intelligence: insights from the social sciences. Artif Intell. 2019;267:1–38. https://doi.org/10.1016/j.artint.2018.07.007

12. Lai V, Tan C. On human predictions with explanations and predictions of machine learning models: a case study on deception detection. In: FAT* ’19. Proceedings of the Conference on Fairness, Accountability, and Transparency; 2019 Jan 29–31; Atlanta, GA. Association for Computing Machinery; c2019. p. 29–38. https://doi.org/10.1145/3287560.3287590

13. Kaur H et al. Interpreting interpretability: understanding data scientists’ use of interpretability tools for machine learning. In: CHI ’20. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; 2020 Apr 25–30; Honolulu, HI. Association for Computing Machinery; c2020. p. 1–14. https://doi.org/10.1145/3313831.3376219

14. Rey GD. A review of research and a meta-analysis of the seductive detail effect. Educ Res Rev. 2012;7(3):216–237. https://doi.org/10.1016/j.edurev.2012.05.003

15. Suh A, Hurley I, Smith N, Siu HC. Fewer than 1% of explainable AI papers validate explainability with humans. In: Yamashita N, Evers V, Yatani K, Ding XS, editors. CHI EA ’25: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems; 2025 Apr 26–May 1; New York, NY. Association for Computing Machinery; c2025. p. 1–7. https://doi.org/10.1145/3706599.3719964

16. Chegere M, Falco P, Nieddiu M, Pandolfi L, Stein M. The magic of the game: experimental evidence on sports betting behavior. CSEF Working Paper No. 655. Centre for Studies in Economics and Finance, University of Naples, Italy; 2022 Nov. Revised 2024 Nov 15. https://www.csef.it/WP/wp655.pdf

17. Green B, Chen Y. Disparate interactions: an algorithm-in-the-loop analysis of fairness in risk assessments. In: FAT* ’19: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency; 2019 Jan 29–31; Atlanta, GA. Association for Computing Machinery; c2019. p. 90–99. https://doi.org/10.1145/3287560.3287563

18. Liao QV, Gruen D, Miller S. Questioning the AI: informing design practices for explainable AI user experiences. CHI ’20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems; 2020 Apr 25–30; Honolulu, HI. Association for Computing Machinery; c2020. p. 1–15. https://doi.org/10.1145/3313831.3376590

19. Mellers B et al. Psychological strategies for winning a geopolitical forecasting tournament. Psychol Sci. 2014;25(5):1106–1115. https://doi.org/10.1177/0956797614524255

20. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. JAIR. 2002;16:321–357. https://doi.org/10.1613/jair.953

21. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: von Luxburg U et al., editors. NIPS '17: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA. Curran Associates Inc; c2017. p. 4768–4777. https://doi.org/10.48550/arXiv.1705.07874

List of Symbols, Abbreviations, and Acronyms

AI	Artificial Intelligence
ANOVA	Analysis of Variance
ARL	Army Research Laboratory
DEVCOM	U.S. Army Combat Capabilities Development Command
H	Hypotheses
RMSE	Root Mean Square Error
RQ	Research Question
SHAP	SHapley Additive exPlanations
std	Standard Deviation
UFC	Ultimate Fighting Championship
XAI	Explainable AI

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
1 (PDF)	DEVCOM ARL FCDD RLB CI TECH LIB