Human Oversight in Action: Designing Risk-Aware AI Systems for Employee Incentive Planning

Yiheng An[1,*], Jia Li[2], Fan Yang[3] and Jiawei Zhang[4]

[1]Warrington College of Business, University of Florida, Gainesville, USA
[2]School of Business, Wake Forest University, Winston-Salem, USA
[3]Department of Computer Science, Wake Forest University, Winston-Salem, USA
[4]Alibaba Group, Hangzhou, China

Abstract

AI systems that manage or influence employees are rapidly entering high-stakes organizational workflows, creating new challenges for operationalizing meaningful human oversight. In this paper, we introduce a practical, in-use case of human oversight for a risk-aware AI system that autonomously plans and targets incentives in real-time customer service operations. We present three complementary mechanisms that enable oversight in practice: (1) causal uplift–based targeting that links incentive allocation to estimated incremental impact, (2) post-hoc extraction of context-specific decision rules that expose the system's behavioral logic in interpretable form, and (3) an interactive oversight interface that supports multi-level monitoring, auditing, and selective intervention during deployment. This case illustrates how human oversight can be engineered into employee-facing AI systems while preserving operational autonomy in real-world deployments.

Keywords

Human Oversight, Incentive Management, Decision Support System, Human-Centered Artificial Intelligence

1. Introduction

Employee-facing AI systems are increasingly deployed to support decisions that directly shape workers' efforts, performance, supervision, and economic opportunity [1, 2, 3, 4]. These systems often operate autonomously and at scale, embedding algorithmic decision-making into core managerial processes. As a result, errors, biases, or misaligned objectives can cause systematic, persistent, and difficult-to-detect harms to workers, affecting worker well-being, as well as how work is valued, rewarded, and governed over time. Despite these risks, meaningful human oversight in such AI systems remains insufficiently specified and underexplored in practice.

Recent regulatory and scholarly discussions emphasize that human oversight should not be treated as an end in itself, but as a mechanism for risk mitigation that leverages human–AI collaboration in high-risk decision-making [5, 6, 7]. In particular, emerging perspectives argue that effective oversight depends on human decision makers having appropriate judgment, sufficient understanding of system behavior, and the ability to influence system outcomes [6, 8], while also emphasizing the need to balance operational effectiveness with regulatory compliance and ethical responsibility in real-world AI system design [9]. However, translating these high-level principles into concrete design choices remains an open challenge, given the complexity spanning technical systems, human decision-making, and organizational contexts.

This paper contributes perspectives grounded in industry practice on oversight-enabled system design by investigating an in-use AI system at Alibaba Group that autonomously plans and targets incentives for customer service employees in real-time operations. The system aims to improve customer satisfaction by proactively offering incentive opportunities to human service agents while they handle customer service cases [3]. This operative autonomy introduces several challenges for meaningful human oversight: (i) *System understandability.* System processes must be traceable and explainable

Joint Proceedings of the ACM Intelligent User Interfaces (IUI) Workshops 2026, March 23-26, 2026, Paphos, Cyprus
*This work was completed prior to Yiheng An's affiliation with the University of Florida.
yiheng.an@ufl.edu (Y. An); lijia@wfu.edu (J. Li); yangfan@wfu.edu (F. Yang); linqiu.zjw@taobao.com (J. Zhang)

and not rely on black-box or opaque models; (ii) *Real-time interpretability.* System behavior should be conveyed through accurate, human-interpretable abstractions that provide oversight actors with timely epistemic access to decision contexts without compromising the system's autonomy or performance. (iii) *Monitoring and intervention.* The system should surface potentially risky decision patterns and allow human decision makers to intervene or override decisions in a timely manner.

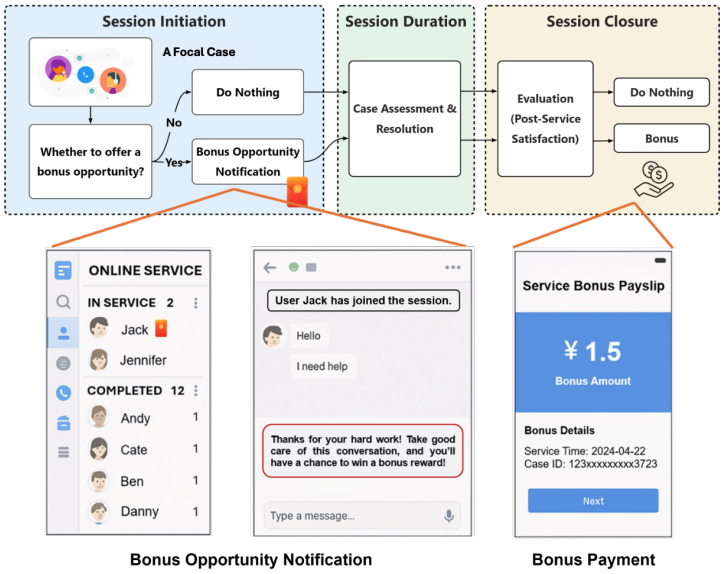
In this paper, we present how thoughtfully designed technical mechanisms and processes can support building oversight-enabled AI systems, exemplified through the incentive-planning practice at Alibaba's customer service center. Specifically, we adopt a causal uplift–based targeting approach that estimates the incremental impact of incentives on employee performance at the service-session level. By prioritizing sessions based on unbiased counterfactual estimates, the system follows a transparent decision logic and mitigates fairness-related risks. Second, we implement a context-aware explainability mechanism that derives human-interpretable and context-specific decision rules from the system's behavior, enabling informed understanding of the decision situation. Third, we design an interactive user interface that supports run-time monitoring from global system-level inspection to local decision-level intervention, enabling human decision makers to explore decision patterns, evaluate and justify anomalies, and selectively override individual decisions when necessary.

Through this case, we demonstrate how causal modeling, post-hoc explainability, and interface design can be intentionally integrated to operationalize human oversight in employee-facing AI systems. By grounding technical design choices in oversight principles, we provide a concrete design blueprint for enabling meaningful human oversight in real-world organizational settings.

2. Operational Setting and Oversight Challenges

To ground our study in a real high-risk employee-facing AI deployment, we partner with Alibaba Group, a Fortune Global 100 technology firm operating a large-scale post-purchase customer service center, and examine an incentive-targeting system for customer service agents. In this setting, customer issues are resolved through live, text-based service sessions between human agents and customers. Each session corresponds to a single customer case and is automatically assigned to an available agent, leaving agents with no discretion over case selection.

Figure 1: Representation of the Customer Service Bonuses Offering Workflow



Notes. When a bonus opportunity is triggered, the agent is notified through two interface cues: a red envelope icon in the navigation bar and a prompt message in the live chat box. The bonus amount is determined after the interaction concludes and is calculated based on the customer's satisfaction rating on a 1–5 scale. If no customer feedback is provided, the agent still receives a bonus under predefined criteria.

Within this environment, the firm plans to introduce proactive incentive interventions to influence employee effort during service interactions. The task of the AI system is to selectively plan and target *bonus opportunities* for human service agents, with the objective of improving customer satisfaction. As illustrated in Figure 1, incentives are presented to agents before the agent-customer interaction begins and remain visible throughout the interaction, functioning as an ex ante signal that can influence how agents allocate attention and effort during the service session.

This operational setting amplifies three core requirements for meaningful human oversight. First, a foundational requirement for oversight is the ability to trace and explain how the system determines when and to whom incentives are offered. When decision logic is embedded in opaque models, the basis for incentive allocation becomes difficult to interpret, creating fairness concerns, and increasing the possibility that decisions reflect unintended or biased correlations, including those associated with protected employee attributes. Meaningful oversight in this context therefore benefits from an inherently explainable targeting approach in which outputs can be traced back to human-understandable factors, allowing humans to interpret how incentive decisions are generated at the system level.

Second, effective oversight requires human decision makers to interpret the system's behavior within the same decision contexts in which incentives are targeted, given the dynamic session-level conditions arising in streaming live customer service interactions. However, raw model outputs or internal parameters are not directly usable for real-time, context-aware reasoning. In this setting, oversight therefore relies on accurate abstractions that summarize context-specific decision criteria and targeting rationale in forms that are cognitively accessible and aligned with operational workflows. These abstractions must provide timely epistemic access to decision contexts while preserving the system's autonomy and performance.

Third, because incentive decisions occur continuously at high frequency across a large stream of service sessions, per-decision review is infeasible. Oversight therefore must operate through monitoring decision patterns across meaningful levels of abstraction (e.g., across time, agents, and session types) to flag risks that are difficult to detect within individual sessions. This creates a need for a user-friendly interface that surfaces these patterns intuitively and supports timely, appropriate intervention, such as drilling down to specific incentive decisions and overriding them when necessary.

3. Oversight-Enabled System Design

To design effective oversight mechanisms in this setting, we integrate three complementary system elements, each aligned with one of the core oversight requirements described above.

3.1. Causal Uplift–Based Targeting for System Understandability

To support traceable and explainable incentive targeting, we adopt a causal uplift–based approach that produces transparent decision logic at the system level. Formally, for each service session i, the system estimates the *conditional average treatment effect* (CATE) of offering an incentive given the observed service context X_i. Let $Y_i(1)$ denote the potential customer satisfaction outcome if an incentive opportunity is offered, and $Y_i(0)$ denote the outcome if no incentive is offered. The session-level causal uplift is defined as:

$$\tau(X_i) = \mathbb{E}\big[Y_i(1) - Y_i(0) \mid X_i\big],$$

where X_i includes contextual features available at the start of the service session, such as session timing, operational conditions, case characteristics, and customer history. This formulation allows the system to reason explicitly about the *expected improvement* attributable to an incentive.

Estimating uplift enables transparent policy-level prioritization. Under a fixed incentive budget, sessions are ranked by their estimated uplift values $\tau(X_i)$, and incentive opportunities are assigned to those with the highest expected improvement. This makes the system's targeting logic explicitly grounded in where incentives are predicted to have an impact.

This uplift-based targeting approach strengthens the system's understandability for oversight. First, it provides a clear and interpretable rationale for why incentives are offered in particular sessions. Second, it allows oversight actors to calibrate the incentive strategy by examining uplift distributions and targeting thresholds. Third, by separating heterogeneous treatment effects of incentives on performance, the approach grounds incentive allocation in effort-responsive signals rather than potentially sensitive or protected features (e.g., gender or age), helping mitigate risks that could otherwise arise from systematically favoring particular agents or session types based solely on outcome predictions.

3.2. Post-Hoc Explainability through Human-Interpretable Decision Rules

While causal uplift modeling supports system-level reasoning, effective oversight also depends on providing epistemic access to how incentive-targeting decisions are enacted under context-specific conditions. To enable run-time interpretability without exposing model internals, we introduce a post-hoc explainability module that derives human-interpretable decision rules from the system's targeting behavior based on contextual features available at decision time.

At a high level, the module approximates the incentive-targeting policy using interpretable surrogate models and extracts a compact set of human-readable rules that summarize how contextual features available at decision time map to targeting outcomes. This approach preserves operational efficiency, as explainability is decoupled from the real-time targeting pipeline, and remains model-agnostic so that oversight mechanisms remain reusable even as the system's decision logic evolve.

The rule extraction involves two conceptual steps. First, we approximate the targeting decisions produced by the AI system using lightweight tree-based surrogate models trained on the same contextual features observable at decision time. This ensures that explanations remain faithful to the decision situation and grounded in information accessible to human reviewers. Second, we apply an optimization-based procedure [10] to derive a parsimonious set of recurring "if–then" rules from the surrogate models. Rather than capturing isolated decision paths, the procedure prioritizes patterns that are consistently recovered across multiple surrogates, filtering out unstable or low-frequency artifacts. The resulting rules reflect dominant policy logic in a stable and interpretable form. Details of the procedures and fidelity evaluation are provided in the Appendix, Section A.

These rules function as context-aware abstractions that allow human decision makers to interpret when incentives are more or less likely to be offered and to assess whether system behavior aligns with organizational intent, risk tolerance, and fairness expectations.

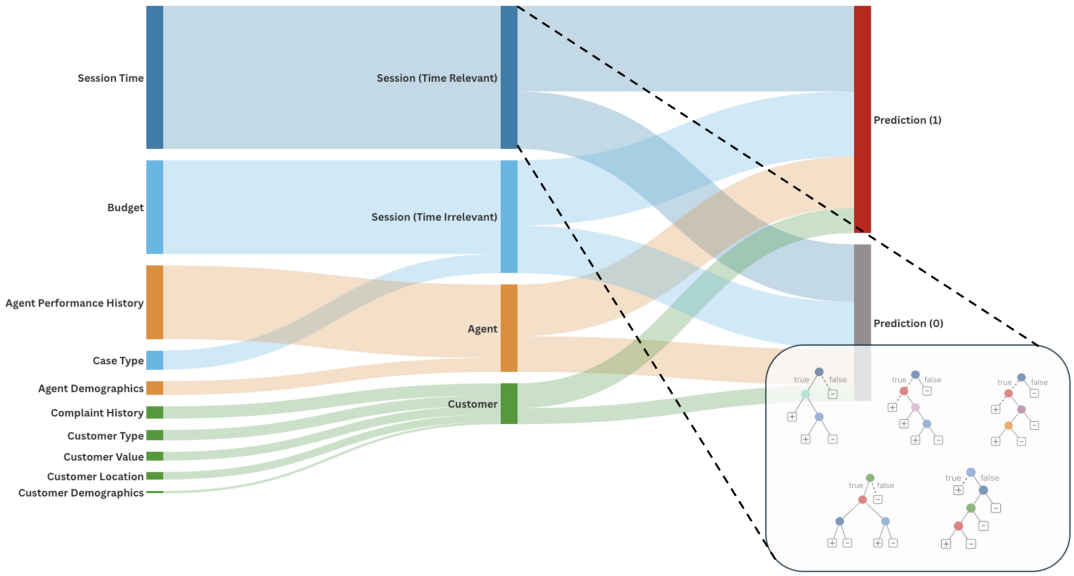
3.3. Oversight-Oriented User Interface for Monitoring and Intervention

To support continuous supervision in a setting where targeting decisions occur at high frequency and scale, we introduce an oversight-oriented user interface designed for deployment-time monitoring and selective intervention.

Figure 2 illustrates the Sankey-style interface that summarizes how groups of contextual features contribute to incentive-targeting outcomes. The widths of the flows correspond to how often features appear in the extracted decision rules, allowing human decision makers to monitor shifts in system reliance patterns and identify potential concentration risks across contextual dimensions.

The interface further supports drill-down from aggregated patterns to representative rule-level logic. As shown in the pop-up view in Figure 2, managers can interactively explore specific flows to inspect the contextual conditions associated with individual incentive decisions. For each selected segment, the interface displays the corresponding extracted rules and key triggering features, providing traceable evidence that supports managerial review and audit of individual incentive allocations. When an allocation is deemed inappropriate or risky, managers can selectively override the decision for that session directly through the interface, ensuring targeted intervention without affecting the system's broader decision logic.

Figure 2: Oversight Interface Visualizing Incentive Decision Logic with Interactive Drill-Down



Notes. The interface presents a Sankey-based visualization constructed from decision rules extracted from the incentive-targeting system. Flow widths reflect how frequently contextual features appear in the rules, providing a system-level view of decision patterns. The interactive drill-down reveals representative rule-level conditions for each segment, enabling monitoring, interpretation, and selective intervention.

4. Discussion and Future Work

This paper presents a practical case of human oversight for an employee-facing AI system that autonomously plans and targets incentives in real-time customer service operations. We show how considerations of system risks motivate design choices that jointly support policy-level reasoning, epistemic access, and continuous monitoring and intervention during deployment.

Our perspectives contribute to ongoing debates on effective human oversight for high-autonomy workplace AI. Prior work emphasizes that oversight effectiveness depends not only on human roles or organizational processes, but critically on technical design features that provide epistemic access and intervention capacity [6]. Consistent with these perspectives, our study illustrates how such features can be concretely realized in a real-world employee management setting.

Specifically, the system integrates three oversight-enabling mechanisms suited to high-frequency customer service environments. First, causal uplift–based targeting provides a transparent rationale for incentive allocation grounded in expected impact rather than outcome prediction alone. Second, post-hoc rule extraction translates complex system behavior into human-interpretable decision logic, improving understandability and supporting evaluative judgment without constraining autonomy. Third, the oversight-oriented interface enables continuous monitoring across multiple levels and supports selective overrides when potentially risky or misaligned patterns emerge. Together, these mechanisms demonstrate how meaningful human oversight of employee-facing AI systems can be achieved through oversight-by-design that balances autonomy, interpretability, and timely human intervention.

Looking forward, this case highlights broader opportunities for designing oversight-ready AI systems in organizational environments where automated decisions directly affect workers. Future work should examine how similar oversight mechanisms generalize across other high-risk employee-facing AI applications, such as scheduling, performance evaluation, and compensation systems, as well as how organizational processes and governance structures interact with technical oversight infrastructures over time. Additional work should also provide greater visibility into human oversight practices by examining who performs oversight, under what organizational constraints, and with what training or incentives. Finally, given the single-case scope of this study, further research should investigate how oversight approaches transfer across domains with differing risk profiles and power dynamics.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-5.2 for grammar and spelling check. After using the tool, the authors reviewed and edited the content as needed and assume full responsibility for the content.

References

[1] M. K. Lee, D. Kusbit, E. Metsky, L. Dabbish, Working with machines: The impact of algorithmic and data-driven management on human workers, in: Proceedings of the 33rd annual ACM conference on human factors in computing systems, 2015, pp. 1603–1612.

[2] M. Raghavan, S. Barocas, J. Kleinberg, K. Levy, Mitigating bias in algorithmic hiring: Evaluating claims and practices, in: Proceedings of the 2020 conference on fairness, accountability, and transparency, 2020, pp. 469–481.

[3] Y. An, J. Li, J. D. Camm, L. Hu, Q. Zhuge, B. Jia, Impact: An inference-driven modeling framework for cost-effective incentive allocation in service operations, in: Proceedings of the 3rd Workshop on Causal Inference and Machine Learning in Practice, the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2025.

[4] M. Dong, L. Brinkmann, O. Sherif, S. Wang, X. Zhang, J.-F. Bonnefon, I. Rahwan, Experimental evidence that ai-managed workers tolerate lower pay without demotivation, arXiv preprint arXiv:2505.21752 (2025).

[5] A. I. Act, Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008,(eu) no 167/2013,(eu) no 168/2013,(eu) 2018/858,(eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu,(eu) 2016/797 and (eu) 2020/1828 (artificial intelligence act), Official Journal of the European Union L 1689 (2024) 2024.

[6] S. Sterz, K. Baum, S. Biewer, H. Hermanns, A. Lauber-Rönsberg, P. Meinel, M. Langer, On the quest for effectiveness in human oversight: Interdisciplinary perspectives, in: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, 2024, pp. 2495–2507.

[7] L. Zhu, Q. Lu, D. Ming, S. U. Lee, C. Wang, Designing meaningful human oversight in ai, Available at SSRN 5501939 (2025).

[8] C. Faas, S. Kerstan, R. Uth, M. Langer, A. M. Feit, Design considerations for human oversight of ai: Insights from co-design workshops and work design theory, arXiv preprint arXiv:2510.19512 (2025).

[9] E. Sadovski, I. Aviv, I. Hadar, Navigating the human-oversight dilemma in ai-based systems, in: 2025 IEEE 33rd International Requirements Engineering Conference Workshops (REW), IEEE, 2025, pp. 454–461.

[10] B. Liu, R. Mazumder, Fire: An optimization approach for fast interpretable rule extraction, in: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023, pp. 1396–1405.

A. The Explainability Module

This section provides additional details of the post-hoc explainability module used to interpret the incentive allocation policy generated by the causal uplift–based targeting system.

A.1. Post-Hoc Rules Generation

Given the uplift–based targeting policy, the explainability module operates in two steps. First, it approximates the policy using interpretable surrogate models. Second, it extracts a compact set of human-readable decision rules from these surrogates that summarize how the policy maps decision-time features to incentive assignments.

(i) Surrogate Policy Modeling. The explainability module takes pre-session context features as input to train interpretable surrogate models that predict the decisions generated by the targeting policy. Specifically, we train a random forest (RF) classifier $s(\cdot)$ to predict binary allocation decisions $\mathbf{y}$ (i.e., offering vs. not offering an bonus opportunity) of the targeting policy using context features X^{pre}. The mapping $\mathbf{y} = s(X^{\text{pre}})$ serves as a surrogate model for the policy, capturing its decision boundaries in a form that is amenable to rule extraction while leaving the policy intact.

(ii) Optimization-Based Rule Extraction. Given the trained RF surrogate model $s(\cdot)$, we adopt the established FIRE framework [10] to extract a representative subset of decision rules from it by solving a regularized optimization problem defined over all leaf nodes of the model:

$$\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{y} - M\mathbf{w}\|_2^2 + h(\mathbf{w}; \lambda_s) + g(\mathbf{w}; \lambda_f), \tag{1}$$

where $\mathbf{y}$ are the targeting policy's decisions, M maps instances to leaf-node predictions, and $\mathbf{w}$ is the vector of rule weights. The quadratic loss $\frac{1}{2}\|\mathbf{y} - M\mathbf{w}\|_2^2$ penalizes deviations between $\mathbf{y}$ and the predictions of the selected set of rules $(M\mathbf{w})$. Minimizing this term ensures the selected rule set faithfully mimics the original policy. The objective also combines a *sparsity penalty* $h(\mathbf{w}; \lambda_s)$ with regularization parameter λ_s to aggressively prune redundant or low-utility rules, and a *fusion penalty* $g(\mathbf{w}; \lambda_f)$ with regularization parameter λ_f to encourage selection of rules sharing common antecedents, resulting in more parsimonious and interpretable rule sets.

(iii) Rule Construction and Representation. Each decision rule corresponds to a leaf node in one of the trees in the RF surrogate model and is defined by the conjunction of split conditions along the path from the root to that leaf. Formally, a rule r_j can be written as

$$r_j(x) = \mathbf{1}\{x \in \mathcal{R}_j\},$$

where $\mathcal{R}_j$ denotes the hyper-rectangle in the feature space induced by the sequence of threshold-based splits leading to leaf j. These rules map pre-session context features to binary allocation outcomes and can be interpreted as *if–then* statements describing regions of the feature space where the targeting policy assigns incentives.

The FIRE optimization problem in (1) assigns a weight w_j to each candidate rule. Rules with nonzero weights in the optimized solution constitute the extracted rule set, while rules with zero weights are pruned. The resulting rule ensemble therefore provides a sparse linear approximation to the targeting policy's allocation decisions, balancing fidelity and interpretability.

(iv) Rule Fusion and Simplification. In addition to sparsity, the fusion penalty $g(\mathbf{w}; \lambda_f)$ encourages selection of rules that share common antecedents within the same decision tree. As shown in [10], this induces grouping of rules along contiguous branches, allowing multiple rules to reuse shared split conditions. As a result, the extracted rules can often be summarized as partial decision trees rather than isolated conditions, substantially reducing the number of unique feature thresholds that must be examined. This property is particularly valuable for stakeholder-facing explanations, as it highlights dominant features and interaction patterns driving incentive allocation.

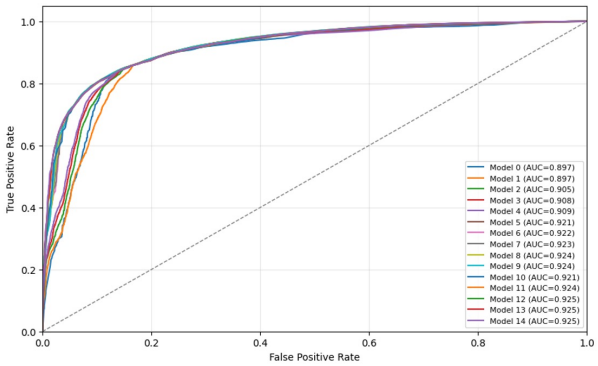
A.2. Post-Hoc Interpretation Fidelity

We first collect the policy's allocation decision logs. The data were then divided into a training set (70%) and a validation set (30%). The training set is used to train surrogate models that approximate the targeting policy's decisions, while the validation set is held out to evaluate how well the surrogate models generalize to unseen sessions.

We begin by training a set of RF classifiers with varying hyperparameter configurations to predict binary policy decisions, using only pre-session features. To ensure that these surrogate models reliably captured the policy decision boundaries, we retained only models that achieved an area under the ROC curve (AUC) of at least 0.9 on the validation set. Figure 3 provides the diagnostic results of the selected models. Across multiple configurations, these RF surrogate models consistently demonstrated high out-of-sample performance, confirming that policy decisions are learnable and can be approximated using observable session-level inputs. This provides strong evidence of interpretation fidelity, as the surrogate models closely replicate the targeting behavior of the causal uplift-based targeting policy.

Then FIRE is applied to each high-performing surrogate to extract the most influential decision rules. We keep only the rules that appear consistently across multiple surrogates, yielding a robust, stable, and compact set of "if–then" rules that summarize the core decision-making logic of the targeting policy. By relying on overlapping rules, the procedure avoids overcommitting to any single surrogate model and ensures that the learned rules are not driven by randomness or model-specific variation.

Figure 3: ROC Curves of Surrogate Policy Models



Notes. Each line shows the ROC curve of a random forest classifier trained to approximate the targeting policy using pre-session features. The consistently high AUC scores indicate that the learned surrogate models are highly accurate in replicating the policy's decisions, supporting the feasibility of extracting interpretable approximations.