December 2025

Optimizing Business Forecasting through Human-AI Decision Fusion: A Theoretical Framework and Simulation Study

Shaohui Wang
Georgia State University, swang83@gsu.edu

Optimizing Business Forecasting Through Human-AI Decision Fusion: A Theoretical Framework and Simulation Study

Short Paper

Shaohui Wang
Georgia State University
Atlanta, GA, USA
swang83@gsu.edu

Abstract

This paper develops a theoretical framework for when and why human–AI ensembles improve decision accuracy. We show that gains arise when at least one agent is better than chance and when complementary errors can be systematically identified, formalized as the agent-complementary information value (ACIV). In multi-round settings, gain from ensembles decompose into conditional VOI terms, highlighting the role of sequential feedback. Building on this theory, we design an online fusion algorithm and test it in simulation to validate our theory.

Keywords: human–AI ensembles, interaction-generated signals, value of information

Introduction

Human-AI collaboration in organizational decision-making has become increasingly popular. Prior research typically frames AI usage as either *automation* (machines replacing human work) or *augmentation* (AI assisting human judgment) (Murray et al., 2021; Raisch & Krakowski, 2021). Automation pertains to the complete replacement of human decision-making with AI algorithms, typically justified in contexts where AI demonstrates clear predictive superiority or equivalent performance at significantly lower costs (Berente et al., 2021; Murray et al., 2021). Augmentation, conversely, emphasizes complementarity, assigning distinct, specialized sub-tasks—often data-driven and computationally intensive—to AI, while reserving more complex, contextual, or higher-order judgment tasks for humans (Jia et al., 2024; Pessach et al., 2020; Wang et al., 2024). Both automation and augmentation inherently rest on the assumption of specialization, where tasks are clearly divisible, and agents (human or AI) can perform specific sub-tasks based on their relative advantages in competence, efficiency, or cost-effectiveness (Baird & Maruping, 2021; Berente et al., 2021).

However, more recent literature propose a third approach: *human-AI ensembles* (Choudhary et al., 2025). In this approach, a human and an AI independently analyze the same problem and then combine their judgments—a "division of labor without specialization". Unlike automation or augmentation, neither agent delegates or replaces the other; instead both contribute to every decision (Choudhary et al., 2025). Choudhary et al. (2025) use the term "Human-AI Ensembles" to define such kind of division of labor and show that "Human-AI Ensembles" can improve accuracy "even when neither human nor AI has a clear advantage". This ensemble view is orthogonal to the usual dichotomy because it allows collaboration even when specialization (i.e., one party is strictly better) is infeasible.

Steyvers et al. (2022) argues that human-AI complementarity exists if the prediction accuracy of a human-AI pair exceeds that of either human-human or AI-AI pairs. However, there is no unified agreement regarding

conditions under which a human-AI ensemble consistently outperforms both human-alone and AI-alone approaches. A meta-analysis by Vaccaro et al. (2024) reveals that most studies fail to recognize humans' unique contributions to human-AI teams, instead concluding that integrating humans tends to degrade team performance when human accuracy is lower than AI. Choudhary et al. (2025), on the contrary, specifies that the effectiveness of human-AI ensembles depends critically on data availability—humans must possess implicit knowledge inaccessible to AI—and predictive diversity, meaning human and AI errors should differ substantially (yet the human does not necessarily need to outperform AI). To reconcile these views, we ask two questions:

RQ1 How can mathematical reasoning rigorously establish the existence of gain from ensembles when human and AI agents jointly participate in decision-making, and what are the boundary conditions for this gain?

RQ2 Based on RQ1, how can an algorithm be designed to achieve the optimal gain from human-AI ensembles?

Two related bodies of literature inform this question. First, the information systems (IS) and behavioral literature investigates factors that help resolve conflicts arising from differing human and AI decisions, aiming to enhance complementarity and overall performance in human-AI teams through conflict mitigation. Such key factors include trust, transparency, and algorithm aversion (Glikson & Woolley, 2020; Jussupow et al., 2024; Li et al., 2023). Studies show that opaque algorithms and unexpected errors can undermine user trust (Glikson & Woolley, 2020). In particular, "algorithm aversion" is well documented: people rapidly lose confidence in AI forecasts after seeing mistakes, even if the algorithm outperforms humans (Dietvorst et al., 2015). Such findings suggest that when AI and human outputs diverge, decision-makers may be biased toward their own judgment or distrust AI suggestions. While allowing users to modify an algorithm's outputs significantly increases their likelihood of using it (Dietvorst et al., 2018).

Second, the computer science (CS) literature explores *algorithmic aggregation* and hybrid models. Classic ensemble learning (e.g., bagging, boosting) combines multiple models' predictions (often via voting or weighted averaging) to improve accuracy (Polikar, 2006). More recently, "learning-to-defer" models train classifiers that can either make a prediction or defer to a human expert (Mozannar & Sontag, 2020). These approaches effectively delegate individual cases to the agent (human or AI) expected to be more accurate. However, they typically result in specialization or sequential engagement (e.g., the AI passes certain cases to the human) rather than joint decision-making on each case.

Our approach differs from both streams. We do *not* specialize tasks to one agent; both human and AI jointly evaluate every decision. We also maintain both agents *concurrently engaged* rather than idling one side. Analytically, we show that a true *fusion* of human and AI signals can outperform either agent alone, even when each is only moderately accurate. In this way, we extend Choudhary et al.(2025) by providing explicit aggregation rules and proofs, showing that continual human-AI collaboration (without specialization) can yield superior decision accuracy.

The structure of this paper is as follows: we first provide a mathematical proof and specify boundary conditions under which human-AI ensembles can enhance performance. Based on these conditions, we develop an online decision-fusion algorithm that approximates the optimal gain attainable in human-AI ensembles. Finally, we validate the effectiveness of our proposed algorithm through a preliminary computational simulation study.

Theoretical Framework of Gain from Human-AI Ensembles

In this section, we develop a formal theoretical definition to analyze when and why human and AI ensembles can yield a performance improvement (which we term *gain from ensembles*) over the better individual decision-maker. We assume a binary decision outcome model (each decision is either correct or incorrect) for each agent on a given input, which simplifies analysis by representing each agent's performance as a Bernoulli random variable. We adopt a binary outcome model primarily to enable tractable analysis. This simplification provides a convenient mathematical foundation (modeling outcomes as Bernoulli variables) but introduces some limitations. For example, it treats all errors as equal, ignoring scenarios with asym-

metric error costs (e.g., false positives vs. false negatives in high-stakes domains). But collapsing outcomes into a binary measure will not preclude fusion strategies that weight decisions by confidence. Despite these trade-offs, the binary model yields a clean analytical insight into human-AI decision fusion.

Symbol	Definition
X_n	Observable feature vector for instance n.
$D_H(X_n),\ D_A(X_n)$	Binary correctness outcome (1 if correct, 0 if incorrect) of the human (H) / AI (A) decision for input X_n.
$D_f(X_n)$	Binary correctness outcome of the fused decision f for input X_n.
$p_H(X_n),\ p_A(X_n)$	Marginal success probabilities: $p_H(X_n) = \Pr(D_H(X_n) = 1)$, $p_A(X_n) = \Pr(D_A(X_n) = 1)$.
$\mathbb{E}[\cdot],\ \Pr(\cdot)$	Expectation and probability operators (with respect to the underlying probability space).
$\Delta(X_n)$	Gain from Ensembles for input X_n: $\Delta(X_n) = \mathbb{E}[D_f(X_n)] - \max\{p_H(X_n),\ p_A(X_n)\}$.
$I(D_H; D_A \mid X_n)$	Conditional mutual information between human and AI decision correctness given X_n, measuring conditional dependence.

Table 1. Key Notation Used in Theoretical Analysis

Defining Gain from Human–AI Ensembles

We define the *instance-level gain from ensembles* $\Delta(X_n)$ as the improvement in expected decision accuracy achieved by joint ensemble signals over the better single-agent signal of human and AI decisions. Formally, for a given input X_n(observable feature vector for instance n) let $D_H(X_n), D_A(X_n) \in \{0, 1\}$ indicate whether the human's and AI's decisions are correct (1) or incorrect (0). Let $D_f(X_n)$ denote the correctness of the fused decision f that combines the human and AI outputs. Then $\Delta(X_n) = \mathbb{E}[D_f(X_n)] - \max\{\mathbb{E}[D_H(X_n)], \mathbb{E}[D_A(X_n)]\}$, where expectations are taken over the randomness in outcomes (conditioned on the input X_n). By definition, $\Delta(X_n) > 0$ indicates that the fusion strategy f outperforms the better of the individual agents on that input, achieving a *gain from ensembles*.

This definition focuses on accuracy (probability of correct decision) as the performance metric. We acknowledge that practical human-AI systems may consider other metrics (speed, cost, fairness, user trust, etc.), but accuracy is a fundamental measure and a natural starting point. Focusing on accuracy allows us to derive theoretical insights which could be extended in future work to other performance dimensions.

Conditions for Positive Gain from Human-AI Ensembles

Theorem 1 (Necessary and Sufficient Condition for Positive instance-level Gain from Ensembles)**.** *Consider the human-AI binary decision fusion for a given feature input $X_n = x$. Let $p_H(x) = \Pr(D_H = 1 \mid X_n = x)$ and $p_A(x) = \Pr(D_A = 1 \mid X_n = x)$ denote the human and AI marginal success (correctness) probabilities at x. Without loss of generality, assume $p_H(x) \geq p_A(x)$. Then a positive fusion gain $\Delta(x) > 0$ is achieved by some fusion rule f if and only if the following two conditions hold:*

(i) ***Non-trivial individual accuracy:*** $\max\{p_H(x),\ p_A(x)\} > 0.5$.

(ii) ***Complementary error coverage, identifiable by an observable discriminator Z:*** *there exists an observable discriminator Z (derived from the agents' outputs at x) such that $\Pr(D_H = 0,\ D_A = 1 \mid X_n = x) > 0$, and, denote the case of disagreement as $D := \{D_H \neq D_A\}$, there exist measurable sets A, B with $\Pr(Z \in A \mid D, X_n = x) > 0$, $\Pr(Z \in B \mid D, X_n = x) > 0$ for which*

$$\Pr(D_H = 1, D_A = 0 \mid Z \in A, D, X_n = x) > \Pr(D_H = 0, D_A = 1 \mid Z \in A, D, X_n = x),$$
$$\Pr(D_H = 1, D_A = 0 \mid Z \in B, D, X_n = x) < \Pr(D_H = 0, D_A = 1 \mid Z \in B, D, X_n = x).$$

Intuitively, among the disagreement cases for this same x, sometimes Z takes values in A (the "trust human" zone), and sometimes in B (the "trust AI" zone). Both happen non-negligibly (not just on a measure-zero edge case).
(In the symmetric case $p_A(x) \geq p_H(x)$, swap the roles of H and A in the display above.)

In words, at least one agent must outperform random guessing; moreover, the event where the other agent corrects the first agent must not only occur but also be identifiable *by some observable discriminator Z on which the better side switches across positive-measure regions inside the disagreement set.*

Proof. **Necessity.** Suppose, for the sake of contradiction, that $\Delta(x) > 0$ even though one of the above conditions fails.

First, if (i) fails, then $p_H(x) \leq 0.5$ and $p_A(x) \leq 0.5$. In this case, both human and AI perform at or below chance level. Any fused decision D_f is some function of D_H and D_A (which themselves carry no information beyond these sub-par performances). Without additional external information, the fused outcome cannot surpass random guessing. In particular, since $\max\{p_H(x), p_A(x)\} \leq 0.5$, we have $\Pr(D_f = 1 \mid X_n = x) \leq 0.5$ for any f, implying $\Delta(x) \leq 0$. This contradicts $\Delta(x) > 0$. Thus condition (i) is necessary.

Next, assume (ii) fails (while (i) holds). Without loss of generality (W.l.o.g.) $p_H(x) \geq p_A(x) > 0.5$ and $\Pr(D_H = 0, D_A = 1 \mid X_n = x) = 0$. This means that *whenever the human is incorrect, the AI is also incorrect* on input x. Equivalently, the set of events where the human errs is a subset of the AI's error events. Formally, $D_H = 0$ implies $D_A = 0$ almost surely given $X_n = x$. In this situation, the AI never "covers" any of the human's mistakes. No matter how we construct the fusion rule, whenever the human is wrong, the AI will be wrong as well, so the fusion will inevitably err on those instances. In fact, the optimal strategy in this case is to always follow the human's decision (since the human is at least as accurate as the AI). The fused accuracy then equals $p_H(x) = \max\{p_H(x), p_A(x)\}$, with $\Delta(x) = 0$. This contradicts the assumption of positive gain. Hence (ii) is also necessary: if the other agent never succeeds when the best agent fails, fusion cannot surpass the best agent. □

Proof. **Sufficiency.** Assume (i) and (ii). Let $C = \{D_H = 1, D_A = 1\}$, $E_H = \{D_H = 1, D_A = 0\}$, $E_A = \{D_H = 0, D_A = 1\}$, and $D = \{D_H \neq D_A\} = E_H \cup E_A$. Then $p_H(x) = \Pr(C \mid x) + \Pr(E_H \mid x)$ and $p_A(x) = \Pr(C \mid x) + \Pr(E_A \mid x)$, hence $\max\{p_H(x), p_A(x)\} = \Pr(C \mid x) + \max\{\Pr(E_H \mid x), \Pr(E_A \mid x)\}$.

Consider the Bayes router f^* that outputs the common label on agreement and, on D, chooses H iff $\Pr(E_H \mid Z, D, x) \geq \Pr(E_A \mid Z, D, x)$ (otherwise chooses A). Then $\Pr(D_{f^*} = 1 \mid x) = \Pr(C \mid x) + \Pr(D \mid x) \cdot \mathbb{E}[\max\{\Pr(E_H \mid Z, D, x), \Pr(E_A \mid Z, D, x)\} \mid D, x]$. By (ii), there exist positive-measure sets A, B under the law of $Z \mid (D, x)$ with $\Pr(E_H \mid Z \in A, D, x) > \Pr(E_A \mid Z \in A, D, x)$ and $\Pr(E_H \mid Z \in B, D, x) < \Pr(E_A \mid Z \in B, D, x)$. Writing $U = \Pr(E_H \mid Z, D, x)$ and $V = \Pr(E_A \mid Z, D, x)$, we have $\Pr(U > V \mid D, x) > 0$ and $\Pr(U < V \mid D, x) > 0$; by convexity of max and the fact that the maximizer is not a.s. constant, $\mathbb{E}[\max\{U, V\} \mid D, x] > \max\{\mathbb{E}[U \mid D, x], \mathbb{E}[V \mid D, x]\} = \max\{\Pr(E_H \mid x)/\Pr(D \mid x), \Pr(E_A \mid x)/\Pr(D \mid x)\}$. Substituting back yields $\Pr(D_{f^*} = 1 \mid x) > \Pr(C \mid x) + \max\{\Pr(E_H \mid x), \Pr(E_A \mid x)\} = \max\{p_H(x), p_A(x)\}$, hence $\Delta(x) > 0$. □

Theorem 2 (Lower Bound on Gain from Human-AI Ensembles)**.** *Let D_H (human) and D_A (AI) be binary decision outcomes (correct vs. incorrect) with accuracies $P(D_H = X_n) = p_H$ and $P(D_A = X_n) = p_A$. The accuracy improvement achieved by an optimal fusion of D_H and D_A, denoted $\Delta(X_n)$, is bounded below by*

$$\Delta(X_n) \geq \frac{\min\{\, p_H(1 - p_H),\ p_A(1 - p_A)\,\}}{\ln 2}\, I(D_H; D_A \mid X_n)\,,$$

where $I(D_H; D_A \mid X_n)$ is the conditional mutual information (Cover, 1999; Shannon, 1948) between the human and AI decisions given the true state. In other words, the fusion gain is at least a positive constant times the informational dependence between D_H and D_A.

Proof Sketch. This bound can be proven using Pinsker's inequality (Csiszár & Körner, 2011) combined with an analysis of decision outcome variance; we omit the technical details here due to space limitations.

This theoretical lower bound provides a justification for pursuing human-AI decision fusion in human-AI Ensembles. It formalizes the intuition that conditional mutual information between the human and AI decisions (i.e., $I(D_H; D_A \mid X_n) > 0$) can be translated into a performance improvement.

Discussion of the Gain from Human-AI Ensembles

Our theory formalizes precise conditions for achieving better performance with human-AI ensembles, addressing gaps from previous qualitative research. Beyond requiring that at least one agent beats chance and that their errors are complementary, we show that complementarity must be identifiable by an observable discriminator Z. This yields clear guidance: collect and calibrate signals that can serve as Z and to gate/triage on Z in disagreement. Organizations in healthcare, finance, and business forecasting can use these insights to effectively integrate AI's quantitative strengths with human expertise.

Connection to Prior Literature on Agent-Complementary Information Value (ACIV)

We previously introduced the instance-level fusion gain $\Delta(x)$, which directly illustrates when and why fusion may outperform the better individual agent. Beyond this instance view, we can extend to a global definition. Following Guo et al. (2025), who define the global agent-complementary information value (ACIV) as an expectation over the data-generating distribution, we define the global fusion gain as $\Delta := R_{\text{fusion}} - \max\{R_H, R_A\}$, where R_H and R_A denote the expected payoff of the human-only and AI-only rules, respectively, and R_{fusion} is the payoff under the fused rule.

In their framework, a decision problem is (π, S), with π the data-generating process and S a payoff function. For any signal V, the Bayesian rational expected payoff is $R_{\pi,S}(V) = \mathbb{E}_{(v,\omega)\sim\pi}[S(d^r(v), \omega)]$, where $d^r(v) = \arg\max_{d\in D} \mathbb{E}_{\omega\sim\pi(\omega|v)}[S(d,\omega)]$. The information value of V is $\text{IV}_{\pi,S}(V) = R_{\pi,S}(V) - R_{\pi,S}(\varnothing)$. They further define $\text{ACIV}(D_b \mid D_\star) = R_{\pi,S}(D_b \cup D_\star) - R_{\pi,S}(D_\star)$, where D_b is the signal of agent $b \in \{H, A\}$ and $D_\star$ is a baseline information set.

In our setting, let D_H and D_A denote the human and AI signals, and write $R_{\pi,S}(D)$ for the payoff from signal D. Define $D_\star := \arg\max\{R_{\pi,S}(D_H), R_{\pi,S}(D_A)\}$ as the stronger agent and D_{other} as the weaker one. Then the fusion gain is $\Delta = R_{\pi,S}(D_H \cup D_A) - \max\{R_{\pi,S}(D_H), R_{\pi,S}(D_A)\} = \text{ACIV}(D_{\text{other}} \mid D_\star)$. Hence $\Delta > 0 \iff \text{ACIV}(D_{\text{other}} \mid D_\star) > 0$, meaning that positive fusion gain arises precisely when the weaker agent contributes complementary information beyond the stronger agent. Intuitively, Δ captures the incremental value that the weaker party's information contributes when added to the stronger party's decision signal.

Gain from Ensembles as Sequential VOI Decomposition

Follow the above discussion, *gain from ensembles is precisely a composition of conditional VOI terms*, depending on whether the incremental contribution flows "Human → AI" or "AI → Human". If the interaction unfolds over a sequence of feedback signals $U_{1:T}$ (e.g., human belief, AI forecast, Human stick/change belief), then the overall fusion gain admits a chain-rule decomposition into incremental VOI terms: $R(\mathcal{I}_0 \cup U_{1:T}) - R(\mathcal{I}_0) = \sum_{t=1}^{T} VOI(U_t \mid \mathcal{I}_0 \cup U_{1:t-1})$, where $\mathcal{I}_0$ is the baseline information. Thus, gain from ensembles is mathematically equivalent to the accumulated VOI of sequential signals, and algorithms that estimate and route based on conditional VOI can provably realize this decomposition in practice.

Algorithm Design for Achieving Gain from Ensembles

Prior work on human-AI collaboration has emphasized that the value of a teammate is not captured solely by its standalone accuracy, but by how effectively human feedback and AI predictions can be integrated to improve team utility. For example, Wilder et al. (2020) formalize the decision of whether to consult a human as a value-of-information (VOI) problem, showing that the marginal gain from human input can be precisely modeled. Bansal et al. (2019, 2021) further highlight that team performance depends critically on how human users update their judgments after observing AI recommendations, and that improvements in AI accuracy alone may not yield better collaboration unless humans can provide complementary signals. Vaccaro et al. (2024) synthesize this literature, noting that most existing studies restrict attention to single-round interactions, while richer multi-round processes may unlock greater complementarities. Building on

these insights, our intuition for algorithm design is to explicitly model multi-round human-AI exchanges, where each successive piece of feedback—whether from the human after seeing the AI's prediction, or from the AI after incorporating human input—constitutes an additional VOI term. Under standard conditions such as conditional independence of signals or diminishing marginal returns, these recursive VOI contributions guarantee non-decreasing expected utility, and thus, multi-round interactions can systematically improve the accuracy of joint decisions.

1: Input: stream $(p_t, h_t^{(0)}, s_t, y_t)_{t=1}^{T}$; heads f_0, f_1, f_2; router W; hyperparams $a, \gamma, \eta_{\text{pol}}, \lambda_{\text{pol}}$.
2: Init: per-head bin tables $\{\mathcal{B}_k\}_{k=0}^{2}$, output bin table $\mathcal{B}^{\text{out}}$; set $W \leftarrow 0$.
3: for $t = 1, \ldots, T$ **do**
  4: *Collect signals* observe $p_t, h_t^{(0)}$; reveal p_t to human; record $s_t \in \{0, 1\}$.
  5: *Candidate probs* $z_t^{(0)} \leftarrow f_0(\phi^{(0)}(p_t))$; $z_t^{(1)} \leftarrow f_1(\phi^{(1)}(p_t, h_t^{(0)}))$;
    $z_t^{(2)} \leftarrow f_2(\phi^{(2)}(p_t, h_t^{(0)}, s_t))$.
  6: *Pre-update local bias & dual-VOI (context only)*
  7: for $k \in \{0, 1, 2\}$ **do** find bin $j_t^{(k)}$ in $\mathcal{B}_k$ (before update);
    $\hat{q}_{k,j} \leftarrow (\sum p_{k,j})/n_{k,j}$; $\hat{\mu}_{k,j} \leftarrow (\sum y_{k,j} + a)/(n_{k,j} + 2a)$;
    $b_t^{(k)} \leftarrow |\hat{\mu}_{k,j} - \hat{q}_{k,j}|$. **end for**
  8: $\Delta_t^{(1)} \leftarrow b_t^{(0)} - b_t^{(1)}$; $\Delta_t^{(2)} \leftarrow b_t^{(1)} - b_t^{(2)}$.
  9: *Contextual Softmax routing & mixture*
  10: $\psi_t \leftarrow [1,\ p_t,\ h_t^{(0)},\ s_t,\ |p_t - h_t^{(0)}|,\ \Delta_t^{(1)},\ \Delta_t^{(2)},\ b_t^{(0)}, b_t^{(1)}, b_t^{(2)}]$.
  11: $w_t \leftarrow \text{softmax}(W\psi_t)$; $\tilde{q}_t \leftarrow \sum_{k=0}^{2} w_{t,k} z_t^{(k)}$.
  12: *Online U-calibration (publish)*
  13: locate output bin j for $\tilde{q}_t$; $\hat{\mu}_j^{\text{out}} \leftarrow (\sum y_j^{\text{out}} + a)/(N_j + 2a)$; $\beta_j \leftarrow e^{-\gamma N_j}$.
  14: $q_t \leftarrow \beta_j \tilde{q}_t + (1 - \beta_j)\hat{\mu}_j^{\text{out}}$; **publish** q_t.
  15: *Observe label and full updates*
  16: observe y_t; update per-head bins in $\mathcal{B}_k$ with $(z_t^{(k)}, y_t)$ for $k = 0, 1, 2$.
  17: for $k \in \{0, 1, 2\}$ **do** recompute $\hat{q}_{k,j}^{\text{post}}, \hat{\mu}_{k,j}^{\text{post}}$;
    $\ell_t^{(k)} \leftarrow |\hat{\mu}_{k,j}^{\text{post}} - \hat{q}_{k,j}^{\text{post}}|$. **end for**
  18: $L_t^{\text{pol}} \leftarrow \sum_k w_{t,k} \ell_t^{(k)}$; $\nabla W_k \leftarrow (\ell_t^{(k)} - L_t^{\text{pol}}) w_{t,k} \psi_t^{\top} + \lambda_{\text{pol}} W_k$;
    $W_k \leftarrow W_k - \eta_{\text{pol}} \nabla W_k$.
  19: one SGD step on each head f_k (cross-entropy with $(z_t^{(k)}, y_t)$).
  20: update output bins $\mathcal{B}^{\text{out}}$ with $(\tilde{q}_t, y_t)$; cache $b_t^{(k)}, \Delta_t^{(1)}, \Delta_t^{(2)}$.
21: end for
22: Output: published probabilities $\{q_t\}_{t=1}^{T}$; learned (W, f_0, f_1, f_2).

Notation. p_t: AI's predicted probability at round t. $h_t^{(0)}$: human's initial belief before seeing AI. $s_t \in \{0, 1\}$: stick/change feedback (1=stick, 0=change). $y_t \in \{0, 1\}$: revealed ground truth. $z_t^{(0)}, z_t^{(1)}, z_t^{(2)}$: candidate forecasts (AI only; AI+human; AI+human+stick/change). $w_{t,k}$: softmax router weights for candidate k. $\tilde{q}_t$: router mixture, $\sum_k w_{t,k} z_t^{(k)}$. q_t: final published probability after U-calibration. ψ_t: context vector including signals, VOI, and bias estimates. $\Delta_t^{(1)}, \Delta_t^{(2)}$: dual VOI terms comparing bias reduction across candidates. $b_t^{(k)}$: local calibration bias estimate for candidate k. $\ell_t^{(k)}$: U-cal aligned loss of candidate k, used to update router.

Figure 1. Heuristic Algorithm for Achieving Gain from Ensembles

Krause and Guestrin (2012) proved that, within graphical models, information gain is a monotone and submodular set function, which guarantees that marginal VOI is always non-negative and typically exhibits diminishing returns. Inspired by their theory, we can readily posit that the recursive interaction is expected to yield positive improvement under the following heuristic conditions: (i) **Informative feedback**: each signal U_t retains positive mutual information with the outcome, $I(Y; U_t \mid \mathcal{I}_0 \cup U_{1:t-1}) > 0$. In practice, this may not always hold; it is sufficient that the signals chosen for interaction exhibit strictly positive VOI. Empirically, this can be validated by measuring the reduction in proper scoring rule loss (e.g., log-loss or Brier score) after incorporating U_t. (ii) **Non-redundancy**: successive signals are not deterministic functions of earlier ones, $H(U_t \mid \mathcal{I}_0 \cup U_{1:t-1}) > 0$. While this condition is not sufficient to guarantee information gain, it provides a testable prerequisite: redundancy can be diagnosed via conditional entropy or feature importance analyses. (iii) **Bounded noise**: feedback must exceed random guessing, e.g., AUC > 0.5, It can be empirically verified by evaluating whether the expected improvement in proper scoring metrics is positive. (iv) **Diminishing but non-negative returns**: marginal VOI remains non-negative under proper scoring

rules, $\mathrm{VOI}(U_t \mid \mathcal{I}_0 \cup U_{1:t-1}) \geq 0$, with equality iff $U_t \perp Y \mid (\mathcal{I}_0 \cup U_{1:t-1})$. In realistic settings, marginal VOI often exhibits diminishing returns, especially under conditional independence of signals given Y. This motivates using heuristics such as stopping rules when estimated marginal VOI falls below a threshold.

Taken together, these conditions provide heuristic guidance: multi-round human–AI interactions are expected to yield monotone improvements in decision quality, provided that feedback is informative, non-redundant, and of sufficient quality. Although some conditions (e.g., strict informativeness or diminishing returns) may only approximately hold in practice, they can be empirically checked using proper scoring metrics and conditional independence tests.

Building on the above discussion, our algorithm design aims to achieve positive gain from ensembles by enabling two rounds of interaction between the human and the AI. Each round contributes additional value of information, and the cumulative effect can be leveraged to optimize the final decision.

Our algorithm integrates three complementary principles to optimize human-AI decision making. First, we extract the VOI from human signals—both the initial belief and the post-AI revision—so that delegation and fusion are justified by information-theoretic gains rather than task-specific utilities (Guo et al., 2025). Second, we ensure forecasts remain robust for any downstream use by enforcing *U-calibration*, which controls external regret across all proper scoring rules and yields sublinear $\tilde{O}(\sqrt{T})$ guarantees (Kleinberg et al., 2023). Third, we minimize *policy regret* by training a contextual softmax router to learn when to rely on human versus AI, addressing the well-documented risk that naïve combinations can underperform either individual (Bansal et al., 2021; Wilder et al., 2020). Together, these layers provide utility-agnostic gains, enabling human-AI teams to outperform solo agents in heterogeneous settings. See Figure 1 for the pseudocode of this algorithm.

A Preliminary Simulation Study

To evaluate the proposed *online* fusion algorithm (context-aware softmax router with dual-VOI and online U-calibration), we conducted a next-week stock price forecasting simulation on 15 randomly selected NASDAQ-listed stocks. To ensure sufficient news coverage, we selected only firms with market capitalization exceeding $10 billion. In our simulation, we did not recruit human participants; instead, we used an LLM as the human proxy. The proxy for *human* is a local LLaMA-3:8b model that reads news articles for the focal firm and outputs an initial prediction about the next week stock price movement of this focal firm. The proxy for *AI* is a random-forest classifier trained on standard market features (price/volume momentum and volatility statistics) to output $P(\text{up next week})$. The entire protocol is *online*: each week we predict first and only then observe the outcome—no static train/test split. A more stringent test would recruit human forecasters; however, our objective here is to validate the fusion algorithm itself. As long as the two predictors draw on distinct information channel (unstructured news vs. structured market data) and exhibit non-trivial error diversity, the resulting interaction signals and gain from ensembles do not hinge on whether the human source is a person or a proxy. Using an LLM proxy also removes uncontrolled variability and cost while preserving the key property we study. We therefore view this setup as a test that does not compromise generality. Controlled human-subjects experiments and field deployments are left for future work.

For each stock we produced weekly up/down predictions over ≈270 weeks (2015–2020). In total, the LLM processed 11,462 news articles, averaging 3.72 articles per firm per week. Market data and firm-level news were sourced from *FNSPID* (Dong et al., 2024). We use the 2015–2020 window to ensure that every firm included in modeling has continuous, week-by-week news coverage across the entire period. In addition, we hide firm identifiers and the original publication timestamps so the LLM cannot leverage prior knowledge or event timing when forming its decisions.

After generating both predictions, we reveal the AI's prediction to the LLM and record whether the LLM *sticks* to or *revises* its view. This stick/flip signal is treated as additional feedback. Our algorithm forms three candidate signals—(i) AI-only, (ii) AI+human initial, and (iii) AI+human+stick/flip—and uses a *contextual softmax router* (with dual VOI features) to learn mixture weights online; the mixed output is then passed through *online U-calibration* to ensure reliability.

Across all 15 firm-level experiments, our fused decision consistently achieved the best performance. We compare our algorithm's fused decisions against *AI-only*, *human-only* (pre- and post-exposure), and standard aggregation rules: weighted voting (Kuncheva & Rodríguez, 2014) and time decay (Kolter & Maloof, 2007). Relative to AI-only, our fusion strategy delivers a 1.42% relative accuracy improvement (p = 0.000053); relative to weighted voting, a 1.48% improvement (p = 0.000139); relative to time decay, a 2.17% improvement (p = 0.000221); relative to human-only pre-exposure, a 57.62% improvement (p < 0.000001); and relative to human-only post-exposure, a 20.69% improvement (p = 0.000002); all differences are statistically significant.

Proposed Future Research Plan and Expected Contribution

This paper develops a rigorous framework for gain from ensembles that treats human-AI interaction as a source of new, decision-relevant signals. Rather than merely combining two static predictions, our algorithm elicits interaction-induced signals (e.g., stick/flip after seeing the AI's estimate) and uses their estimated value of information to adapt routing and weights in real time. We show when interaction-generated signals raise utility beyond either agent.

In future studies, we plan to further validate our algorithm by conducting experiments that systematically examine its effectiveness in human-AI ensembles and to test multi-round interactions between human and AI to identify protocols that maximize VOI and ensemble gains across domains. We will model the resulting signal sequence $S_{1:T}$ and estimate its marginal and cumulative VOI relative to baseline information I_0, testing whether these interactions increase $I(Y; S_{1:T} \mid I_0)$ and thereby raise the attainable *gain from ensembles*. This line of work will yield actionable guidance for training and upskilling humans to work effectively with AI in human-AI ensembles and designing interaction protocols that maximize the gains of human-AI ensembles.

Acknowledgements

This research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-23-2-0224. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Laboratory or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Baird, A., & Maruping, L. M. (2021). The next generation of research on is use: A theoretical framework of delegation to and from agentic is artifacts. *MIS quarterly*, *45*(1). https://doi.org/10.25300/MISQ/2021/15882

Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., & Horvitz, E. (2019). Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. *Proceedings of the AAAI conference on artificial intelligence*, *33*(01), 2429–2437. https://doi.org/10.1609/aaai.v33i01.33012429

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., & Weld, D. S. (2021). Does the whole exceed its parts? the effect of ai explanations on complementary team performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3411764.344571

Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Managing artificial intelligence. *MIS quarterly*, *45*(3). https://doi.org/10.25300/MISQ/2021/16274

Choudhary, V., Marchetti, A., Shrestha, Y. R., & Puranam, P. (2025). Human-ai ensembles: When can they work? *Journal of Management*, *51*(2), 536–569. https://doi.org/10.1177/01492063231194968

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons. https://doi.org/10.1002/047174882X

Csiszár, I., & Körner, J. (2011). *Information theory: Coding theorems for discrete memoryless systems*. Cambridge University Press. https://doi.org/10.1017/CBO9780511921889

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General*, *144*(1), 114. https://doi.org/10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management science*, *64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Dong, Z., Fan, X., & Peng, Z. (2024). Fnspid: A comprehensive financial news dataset in time series. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4918–4927. https://doi.org/10.1145/3637528.3671629

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of management annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Guo, Z., Wu, Y., Hartline, J., & Hullman, J. (2025). The value of information in human-ai decision-making. *arXiv preprint arXiv:2502.06152*. https://doi.org/10.48550/arXiv.2502.06152

Jia, N., Luo, X., Fang, Z., & Liao, C. (2024). When and how artificial intelligence augments employee creativity. *Academy of Management Journal*, *67*(1), 5–32. https://doi.org/10.5465/amj.2022.0426

Jussupow, E., Benbasat, I., & Heinzl, A. (2024). An integrative perspective on algorithm aversion and appreciation in decision-making. *MIS Quarterly*, *48*(4). https://doi.org/10.25300/MISQ/2024/18512

Kleinberg, B., Leme, R. P., Schneider, J., & Teng, Y. (2023). U-calibration: Forecasting for an unknown agent. *The Thirty Sixth Annual Conference on Learning Theory*, 5143–5145. http://proceedings.mlr.press/v195/kleinberg23a.html

Kolter, J. Z., & Maloof, M. A. (2007). Dynamic weighted majority: An ensemble method for drifting concepts. *The Journal of Machine Learning Research*, *8*, 2755–2790. https://doi.org/10.5555/1314498.1390333

Krause, A., & Guestrin, C. (2012). Near-optimal nonmyopic value of information in graphical models. *arXiv preprint arXiv:1207.1394*. https://doi.org/10.48550/arXiv.1207.1394

Kuncheva, L. I., & Rodríguez, J. J. (2014). A weighted voting framework for classifiers ensembles. *Knowledge and information systems*, *38*, 259–275. https://doi.org/10.1007/s10115-012-0586-6

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2023). Trustworthy ai: From principles to practices. *ACM Computing Surveys*, *55*(9), 1–46. https://doi.org/10.1145/3555803

Mozannar, H., & Sontag, D. (2020). Consistent estimators for learning to defer to an expert. *Proceedings of the 37th International Conference on Machine Learning (ICML)*, *119*, 7076–7087. https://proceedings.mlr.press/v119/mozannar20b.html

Murray, A., Rhymer, J., & Sirmon, D. G. (2021). Humans and technology: Forms of conjoined agency in organizations. *Academy of Management Review*, *46*(3), 552–571. https://doi.org/10.5465/amr.2019.0186

Pessach, D., Singer, G., Avrahami, D., Ben-Gal, H. C., Shmueli, E., & Ben-Gal, I. (2020). Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming. *Decision support systems*, *134*, 113290. https://doi.org/10.1016/j.dss.2020.113290

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, *6*(3), 21–45. https://doi.org/10.1109/MCAS.2006.1688199

Raisch, S., & Krakowski, S. (2021). Artificial Intelligence and Management: The Automation–Augmentation Paradox. *Academy of Management Review*, *46*(1), 192–210. https://doi.org/10.5465/amr.2018.0072

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Steyvers, M., Tejeda, H., Kerrigan, G., & Smyth, P. (2022). Bayesian modeling of human–AI complementarity. *Proceedings of the National Academy of Sciences*, *119*(11), e2111547119. https://doi.org/10.1073/pnas.2111547119

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and ai are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 1–11. https://doi.org/10.1287/mnsc.2021.00588

Wang, W., Gao, G., & Agarwal, R. (2024). Friend or foe? teaming between artificial intelligence and workers with variation in experience. *Management Science*, *70*(9), 5753–5775. https://doi.org/10.1287/mnsc.2021.00588

Wilder, B., Horvitz, E., & Kamar, E. (2020). Learning to complement humans. *arXiv preprint arXiv:2005.00582*. https://doi.org/10.48550/arXiv.2005.00582