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

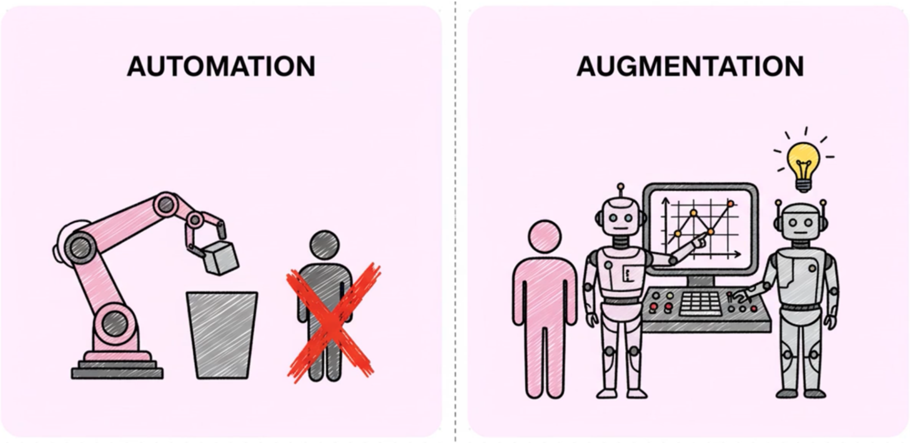
Shaohui Wang

Present at ICIS 2025, Nashville
December 16, 2025

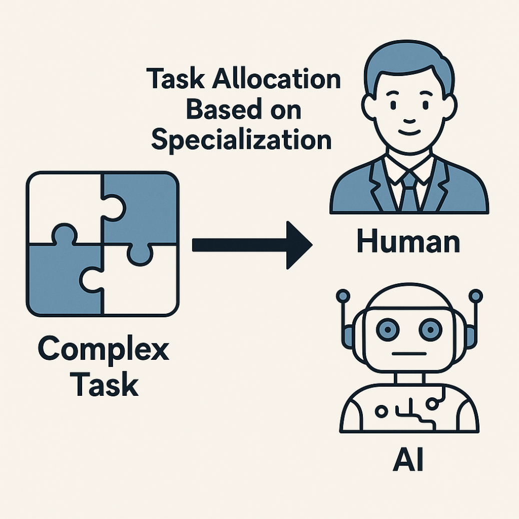
Georgia State University

Replace or Assist Debate

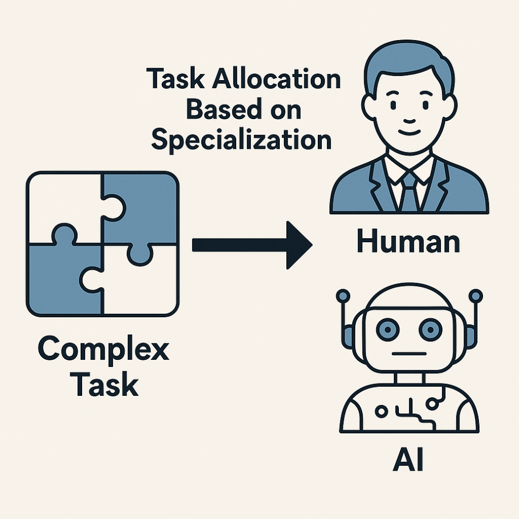
Human-AI collaboration is often framed as **automation** (AI replacing humans) [Berente et al., 2021, Murray et al., 2021] or **augmentation** (AI assisting humans)[Jia et al., 2024, Pessach et al., 2020, Wang et al., 2024], both relying on task specialization [Baird and Maruping, 2021, Berente et al., 2021].



Human-AI Ensembles



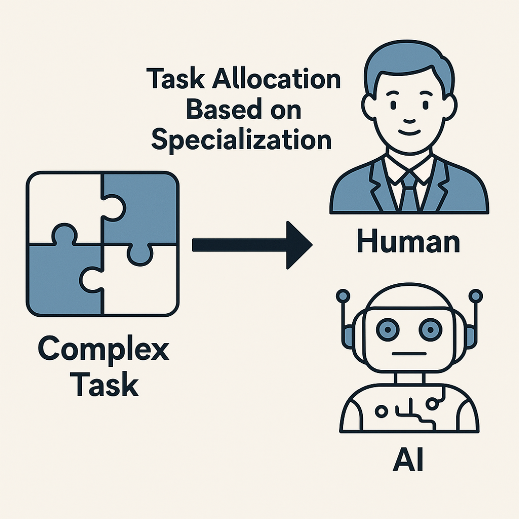
Human-AI Ensembles



What if tasks can't be divided?

- When neither the human nor the AI has a clear advantage in performing a specific sub-task.
- When an organizational task cannot be decomposed into distinct sub-tasks for specialization, an alternative approach is **Human-AI Ensembling** [Choudhary et al., 2025].

Human-AI Ensembles



What if tasks can't be divided?

- When neither the human nor the AI has a clear advantage in performing a specific sub-task.
- When an organizational task cannot be decomposed into distinct sub-tasks for specialization, an alternative approach is **Human-AI Ensembling** [Choudhary et al., 2025].

When Do Humans and AI Need to Handle the Same Task Together?

- High-Stakes Decision Domain [Bansal et al., 2019]
- Complementarities Between Human and AI [Fügener et al., 2021]
- Information Asymmetry Between Human and AI[Hemmer and et al., 2024]
- Capability Asymmetry Between Human and AI [Hemmer and et al., 2024, Steyvers et al.]

Performance Gain from Human-AI Ensembles

What We Know About Human-AI Ensembles

- Involves both agents independently analyzing the same problem and then combining their judgments—a ”division of labor without specialization” [Choudhary and et al., 2023]
- The effectiveness of ensembles critically depends on data availability (humans possessing implicit knowledge) and predictive diversity (different error patterns between human and AI) [Choudhary and et al., 2023]

What We Don’t Know

- RQ1: When do Human-AI Ensembles lead to better performance?
- RQ2: How can we design an algorithm to get the best performance from Human-AI Ensembles?

Theorem 1 — Condition of Performance Gain

Setup. For a given instance x, let the human and AI have accuracies $p_H(x) = \Pr(D_H = Y \mid x)$ and $p_A(x) = \Pr(D_A = Y \mid x)$. Let D_f be the fused decision. The instance-level gain from ensembling is

$$\Delta(x) = \mathbb{E}[\mathbf{1}\{D_f = Y\} \mid x] - \max\{p_H(x), p_A(x)\}.$$

Theorem 1 (Necessary and Sufficient Condition for Positive instance-level Gain from Ensembles)

$\Delta(x) > 0$ **if and only if** two conditions hold:

1. **Non-trivial accuracy:** At least one agent performs better than random guessing, i.e., $\max\{p_H(x), p_A(x)\} > 0.5$.
2. **Identifiable complementary errors:** There exists an observable discriminator Z such that, on the *disagreement set* $\mathcal{S}(x) = \{\omega : D_H(\omega, x) \neq D_A(\omega, x)\}$, there are regions of positive measure where the human is more accurate and regions where the AI is more accurate. In other words, the better performer switches across subregions of $\mathcal{S}(x)$ in a way that can be identified by Z.

Intuition

A strict ensemble gain occurs only when:

- **At least one agent is reliable** — both cannot be random guessers.
- **Their errors are complementary and recognizable** — there must exist an observable signal Z that tells us when to trust the human and when to trust the AI.

Without (i), no gain is possible beyond 0.5 accuracy. Without (ii), the complementary information cannot be exploited, and the best strategy reduces to always choosing the single better agent.

Proof Sketch

Sufficiency. Define a Bayesian router $r^*(x, z) = \arg\max\{\Pr(Y = 1 \mid \text{choose H}, x, z), \Pr(Y = 1 \mid \text{choose A}, x, z)\}$. Because Z identifies regions where the better agent switches, the expected accuracy of r^* strictly exceeds $\max\{p_H(x), p_A(x)\}$, yielding $\Delta(x) > 0$.
Necessity. If condition (i) fails, accuracy cannot exceed 0.5. If condition (ii) fails, either there is no disagreement or the better agent cannot be identified from any observable signal; then any measurable fusion rule collapses to always choosing one agent, giving $\Delta(x) = 0$.

Note. Z can represent any observable cue—features.

Connection to ACIV and Sequential VOI Decomposition

1. Connection to Agent-Complementary Information Value (ACIV)

- The fusion gain Δ measures how much the fused decision outperforms the better individual agent.
- Following Guo et al. [2025], the global agent-complementary information value (ACIV) is:

$$\mathrm{ACIV}(D_b \mid D_*) = R_{\pi,S}(D_b \cup D_*) - R_{\pi,S}(D_*),$$

  where D_* is the stronger agent's signal and D_b is the weaker agent's signal.
- In our setting, $\Delta = R_{\pi,S}(D_H \cup D_A) - \max\{R_{\pi,S}(D_H), R_{\pi,S}(D_A)\} = \mathrm{ACIV}(D_{\mathrm{other}} \mid D_*)$.
- Hence, $\Delta > 0$ if and only if the weaker agent provides complementary information beyond the stronger one.

2. Gain from Ensembles as Sequential VOI Decomposition

- The total fusion gain equals the accumulated **Value of Information (VOI)** across sequential feedback signals:

$$R(I_0 \cup U_{1:T}) - R(I_0) = \sum_{t=1}^{T} VOI(U_t \mid I_0 \cup U_{1:t-1}),$$

  where U_t represents each human or AI feedback signal.
- This means ensemble gain can be viewed as the sum of incremental VOI terms—each round of interaction adds measurable value.

Algorithm Design for Human–AI Ensemble Decision Fusion (I)

Goal: To construct an online algorithm that fuses human and AI decisions so that the ensemble achieves a higher accuracy than either agent alone.

Core Idea

Treat each human–AI interaction as a sequence of information updates:

$$I_0 \to U_1 \to U_2 \to \dots \to U_T,$$

where each U_t is a feedback or correction signal (e.g., human revision after seeing AI's prediction). The algorithm maximizes total **Value of Information (VOI)**:

$$R(I_0 \cup U_{1:T}) - R(I_0) = \sum_{t=1}^{T} VOI(U_t \mid I_0 \cup U_{1:t-1}).$$

Main Components

1. **VOI Extraction:** Quantifies how much each new signal (human or AI) improves expected decision quality.
2. **U-Calibration:** Ensures forecast probabilities remain reliable across unknown downstream scoring rules; minimizes external regret.

Algorithm Design for Human–AI Ensemble Decision Fusion (II)

Algorithm Workflow

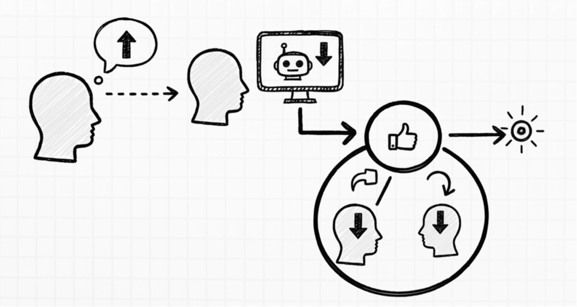
1. **Initialization:** Both human and AI provide independent probability forecasts on a task (e.g., stock increase or not).
2. **Interaction:** The human observes the AI's prediction and decides whether to *stick* or *flip*. This produces three candidate signals:
   - AI-only prediction
   - AI + Human initial belief
   - AI + Human + Stick/Flip response
3. **Routing and Fusion:** A contextual softmax router assigns adaptive weights to each candidate signal based on contextual features and VOI estimates:

$$w_i = \frac{\exp(\theta^\top z_i)}{\sum_j \exp(\theta^\top z_j)}, \quad P_f = \sum_i w_i P_i.$$

   The fused probability P_f is then recalibrated using U-calibration to ensure reliability.
4. **Online Update:** After observing the true outcome, the router and calibration parameters are updated to minimize cumulative regret over time.

Simulation Study: Financial Forecasting with Human–AI Ensembles

Goal: To test whether the proposed Human–AI ensemble algorithm can improve forecasting accuracy in a real-world, high-stakes decision domain.



Forecasting Setting

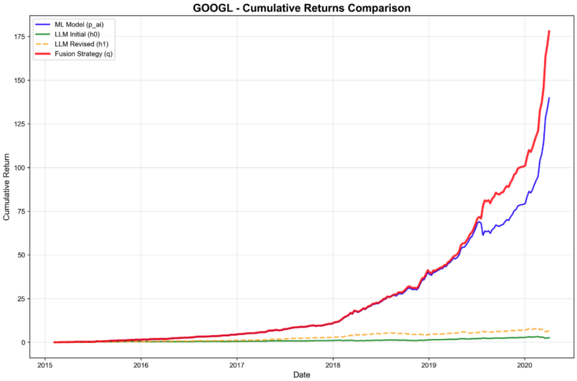
- **Task:** Predict next-week stock price movement (up or down) for 15 NASDAQ-listed firms, 2015–2020.
- **Human Proxy:** A local LLaMA-3 8B model reading company news and generating weekly predictions. In total, the LLM processed 11,462 news articles, averaging 3.72 articles per firm per week.
- **AI Proxy:** A Random Forest classifier trained on financial indicators and historical prices.
- **Interaction:** The human model sees the AI's forecast and either *sticks* or *flips* its decision.

Evaluation

- Compare AI-only, Human-only, and Fused decisions.
- Measure accuracy improvement and statistical significance.

Example: GOOGL Forecasting — Fusion Strategy vs. ML Model

Context: The figure shows cumulative returns from 2015–2020 using four strategies on GOOGL:



Key Observations

- The **ML Model** (blue) performs well on average but fails in certain periods of high uncertainty.
- The **Fusion Strategy** (red) consistently outperforms, even during ML downturns, by leveraging additional cues from the **LLM's revised judgment**.
- This shows that when Theorem 1's conditions hold: (1) at least one model is better than chance, and; (2) their errors are complementary and identifiable, the fusion can exceed both individual agents.

Interpretation

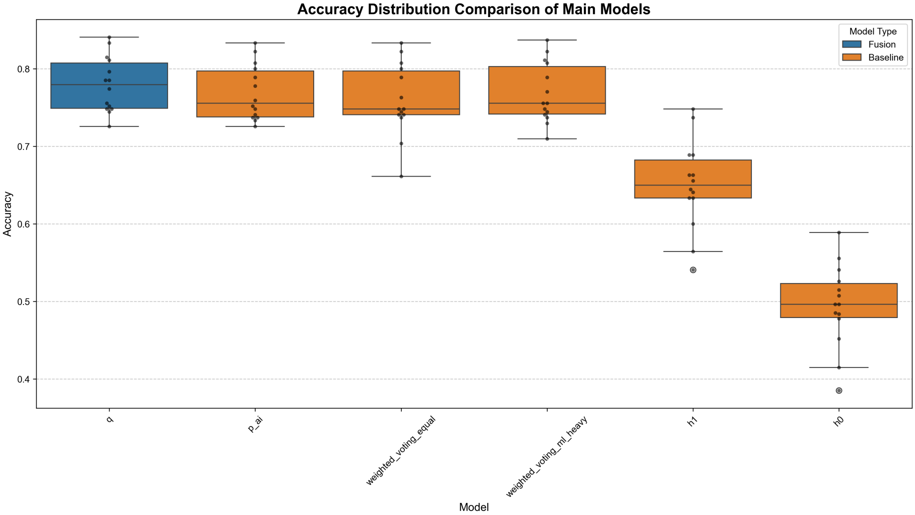
Even though the ML model is intentionally stronger overall (reflecting realistic AI dominance), the fusion algorithm can still detect and exploit moments when the LLM provides valuable corrective signals, yielding a strictly higher cumulative return.

Fusion vs. Baselines — Accuracy Gain with Significance

Baseline Model	**Mean Improvement**	**Rel. Improve. (%)**	**t-test *p*-value**
h0	+0.2850	+57.62	0.000000***
random	+0.2640	+51.20	0.000000***
h1	+0.1295	+19.92	0.000002***
moving_average_5	+0.0221	+2.91	0.000316***
weighted_voting_equal	+0.0197	+2.59	0.002314**
time_decay_fast	+0.0196	+2.58	0.000277***
moving_average_10	+0.0177	+2.33	0.000714***
time_decay_slow	+0.0166	+2.17	0.000221***
weighted_voting_ml_heavy	+0.0112	+1.46	0.000139***
p_ai	+0.0108	+1.41	0.000053***

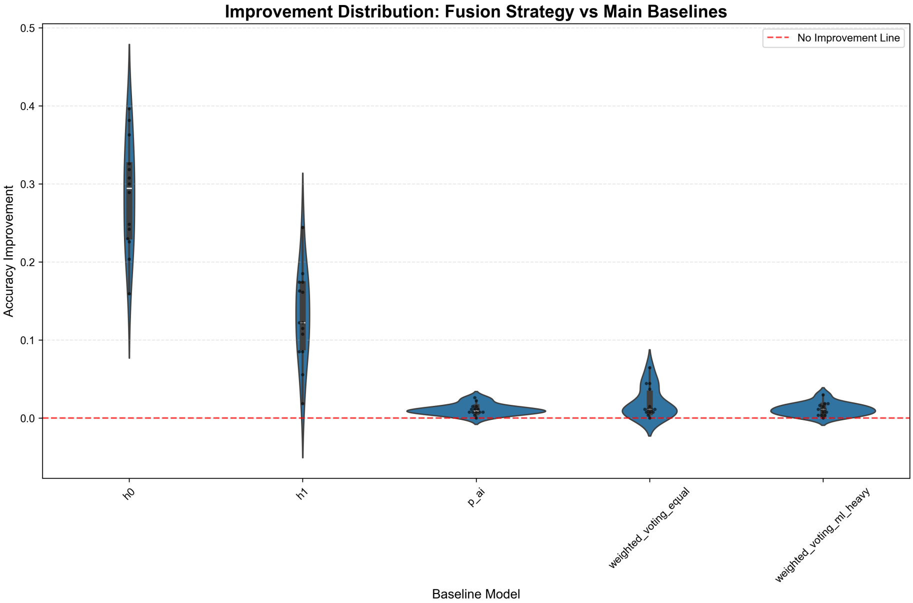
Notes: Mean/relative improvements are for the **Fusion** strategy over each baseline. Significance: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Accuracy Comparison Across Models



The Fusion model (q) achieves the highest and most stable accuracy. Baselines such as AI-only (p_{ai}) and LLM-only (h_0, h_1) perform worse on average.

Accuracy Comparison Across Models



Future Research Plan and Conclusion

Future Research Directions

- **Multi-round Human–AI Interaction:** Extend the current single-feedback design to multi-stage settings, where humans and AI iteratively update beliefs.
- **Behavioral Considerations:** I plan to investigate whether placing humans in an AI-dominant decision environment—where the algorithm has the final say—induces social loafing, causing humans to contribute weaker or less informative signals, thereby reducing the algorithm's ability to learn meaningful Value of Information (VOI). Currently, I am conducting experiments to test whether, under the condition of "When AI is the Boss," human disengagement limits the expected informational gain from collaboration.

Conclusion

- The paper introduces a theoretically grounded algorithm for **Human–AI ensemble decision fusion**.
- The study demonstrates that even when AI dominates in average accuracy, **human feedback still adds measurable value** under identifiable complementarity.

Takeaway: Human–AI complementarity is not about who is smarter, but about designing mechanisms that **detect and utilize when the other side sees what you don't**.

References I

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

Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 2429–2437, 2019. doi: 10.1609/aaai.v33i01.33012429.

Nicholas Berente, Bin Gu, Jan Recker, and Radhika Santhanam. Managing artificial intelligence. *MIS quarterly*, 45(3), 2021. doi: 10.25300/MISQ/2021/16274.

Vivek Choudhary and et al. Human and ai ensembles: When can they work? *Journal of Management*, 2023. Forthcoming.

Vivek Choudhary, Alessandro Marchetti, Yash Raj Shrestha, and Phanish Puranam. Human-ai ensembles: When can they work? *Journal of Management*, 51(2):536–569, 2025. doi: 10.1177/01492063231194968.

References II

Andreas Fügener, Joern Grahl, Alok Gupta, and Wolfgang Ketter. Will humans-in-the-loop become borgs? merits and pitfalls of working with ai. *MIS Quarterly*, 45(3):1527–1556, 2021. doi: 10.25300/MISQ/2021/16553.

Ziyang Guo, Yifan Wu, Jason Hartline, and Jessica Hullman. The value of information in human-ai decision-making. *arXiv preprint arXiv:2502.06152*, 2025. doi: 10.48550/arXiv.2502.06152.

Patrick Hemmer and et al. Complementarity in human-ai collaboration: Concept, sources, and evidence. *arXiv preprint arXiv:2401.10987*, 2024.

Nan Jia, Xueming Luo, Zheng Fang, and Chengcheng Liao. When and how artificial intelligence augments employee creativity. *Academy of Management Journal*, 67(1):5–32, 2024. doi: 10.5465/amj.2022.0426.

Alex Murray, JEN Rhymer, and David G Sirmon. Humans and technology: Forms of conjoined agency in organizations. *Academy of Management Review*, 46(3):552–571, 2021. doi: 10.5465/amr.2019.0186.

References III

Dana Pessach, Gonen Singer, Dan Avrahami, Hila Chalutz Ben-Gal, Erez Shmueli, and Irad Ben-Gal. Employees recruitment: A prescriptive analytics approach via machine learning and mathematical programming. *Decision support systems*, 134:113290, 2020. doi: 10.1016/j.dss.2020.113290.

Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. Bayesian modeling of human–AI complementarity. 119(11):e2111547119. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2111547119. URL https://pnas.org/doi/full/10.1073/pnas.2111547119.

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