Unexploited Information Value in Human-AI Collaboration

Ziyang Guo
Department of Computer Science
Northwetsern University
Evanston, IL, 60208
ziyang.guo@northwestern.edu

Yifan Wu
Department of Computer Science
Northwetsern University
Evanston, IL, 60208
yifan.wu@u.northwestern.edu

Jason Hartline
Department of Computer Science
Northwetsern University
Evanston, IL, 60208
hartline@northwestern.edu

Jessica Hullman
Department of Computer Science
Northwetsern University
Evanston, IL, 60208
jhullman@northwestern.edu

Abstract

Humans and AIs are often paired on decision tasks with the expectation of achieving *complementary performance* – where the combination of human and AI outperforms either one alone. However, how to improve performance of a human-AI team is often not clear without knowing more about what particular information and strategies each agent employs. In this paper, we propose a model based in statistically decision theory to analyze human-AI collaboration from the perspective of what information could be used to improve a human or AI decision. We demonstrate our model on a deepfake detection task to investigate seven video-level features by their unexploited value of information. We compare the human alone, AI alone and human-AI team and offer insights on how the AI assistance impacts people's usage of the information and what information that the AI exploits well might be useful for improving human decisions.

1 Introduction

As the performance of artificial intelligence (AI) models makes remarkable advances, workflows in which humans and AIs collaborate have been sought for important decisions in medicine, finance, and other domains. Designing for human involvement is critical. While an AI model can usually make predictions with higher accuracy than the average human when the two use similar information [Ægisdóttir et al., 2006, Grove et al., 2000, Meehl, 1954], in some cases a human must retain final control over the decision for liability reasons. When humans have access to additional information over the AI, there is the potential for a human-AI collaboration to achieve *complementary performance*, i.e., better performance than either the human or AI alone. For example, a physician may have access to additional information that may not be captured in tabular electronic health records or other structured data [Alur et al., 2024]. Others argue that human theory-based causal logic can contribute knowledge that AI data-based predictions can not learn from historic data [Felin and Holweg].

However, evidence supporting complementary performance between humans and AI is limited, with many studies showing that human-AI teams often underperform AI alone in tasks [Buçinca et al., 2020, Bussone et al., 2015, Green and Chen, 2019, Jacobs et al., 2021, Lai and Tan, 2019, Vaccaro and Waldo, 2019, Kononenko, 2001]. Thus, numerous empirical studies attempt to explore design strategies and conditions under which complementary human-AI performance can be achieved. For

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

example, some find that complementary performance is more likely to be obtained when the AI and human have comparable ability [Bansal et al., 2021]. Other studies focus on improving the workflow [Buçinca et al., 2021, Fogliato et al., 2021] and information display [Bussone et al., 2015, Fok and Weld, 2023].

Most analyses of human behavior in human-AI collaboration to date focus on the *performance* of human-AI teams or each individually, without considering the potential for available *information* to improve the decisions. However, knowing how much more effectively decision-relevant information might be *used* by both agents paves the way to opportunities to improve human use of available information, such as through the design of new explanations or human-AI workflows. A decision-theoretic conceptual framework by Guo et al. [2024] upper bounds the possible performance of an AI-human team using the expected score of a rational Bayesian agent faced with deciding between human and AI recommendations. This approaches provides a basis for identifying informational "opportunities" within a decision problem.

In this paper, we present a method for identifying unexploited information value by a human and an AI in a human-AI collaboration. We use the notion of information gain (the marginal value that one piece of information provides over another) to capture the overlap between the information contained in human decisions (or AI predictions) and the contextual information available to the human. A larger information gain identifies a higher unexploited value of information that a contextual signal offers which might improve decisions. We use the Shapley value [Shapley, 1953] to quantify the contribution of each basic element of information to the overall information value contained in the human decisions. We demonstrate our methodology on a deepfake video detection task [Groh et al., 2022]. Through a comaprision between the information gain over human decisions and AI predictions, we find that participants failed to make use of the considerable information value of some signals that the AI exploited effectively, highlighting the potential for improvement in human decisions. We also find that simply displaying the AI predictions did not necessarily help participants improve on their usage of the information, which suggests the need for further improvements such as explanations of the AI's decision rule on those unexploited signals.

2 Model Setup

Information can be considered valuable to a decision-maker to the extent that it is possible in theory to incorporate it in their decisions to improve performance. Our approach analyzes the expected marginal payoff gain from ideal use of additional information over ideal use of the existing information in human decisions in decision tasks. In this section, we define the basis of this approach, including a decision problem and associated information structure, following prior decision-theoretic frameworks for studying decisions from statistical information [Wu et al., 2023, Guo et al., 2024, Hullman et al., 2024]. Then we define how a rational decision maker would act given a signal and a decision problem with an associated information structure. Using the rational decision maker as a tool, we show how to investigate the information encoded in behavioral decisions.

Decision Problem A decision problem consists of three key elements. We illustrate with an example of a weather decision.

• A payoff-relevant state ω from a space $\mathbf{\Omega}$. For example, $\omega \in \mathbf{\Omega} = \{0, 1\} = \{\text{no rain}, \text{rain}\}$.

• A decision d from the decision space $\mathbf{D}$ characterizing the decision-maker (DM)'s choice. For example, $d \in \mathbf{D} = \{0, 1\} = \{\text{not take umbrella}, \text{take umbrella}\}$.

• The payoff function $S : \mathbf{D} \times \mathbf{\Omega} \to \mathbb{R}$, used to assess the quality of a decision given a realization of the state, e.g., $S(d = 0, \omega = 0) = 0, S(d = 0, \omega = 1) = -100, S(d = 1, \omega = 0) = -50, S(d = 1, \omega = 1) = 0$, which punishes the DM for selecting an action that does not match the weather.

Information Model We cast the information available to a DM as a signal defined within an information structure. We use the definition of an information structure in Blackwell et al. [1951]. The information structure has two elements:

• *Signals*. There are n "basic signals" represented as random variables $\Sigma_1, \ldots, \Sigma_n$, from the signal spaces $\mathbf{\Sigma}_1, \ldots, \mathbf{\Sigma}_n$. These represent information obtained by a decision-maker, e.g., $\Sigma_1 = \{\text{cloudy}, \text{not cloudy}\}$, $\Sigma_2 \in \{0, \ldots, 100\}$ for temprature Celsius, etc.

The decision maker observes a signal, which is a combination of the basic signals, represented as a set $V \subseteq 2^{\{\Sigma_1,\ldots,\Sigma_n\}}$. For example, a signal $V = \{\Sigma_1, \Sigma_2\}$ observed by the decision maker might consist of cloudiness Σ_1 and the temperature Σ_2 of the day. Given a signal composed of m basic signals, we write the realization of V as $v = (\sigma_{j_1}, \ldots, \sigma_{j_m})$, where the realizations are sorted by the index of the basic signals and $\sigma_{j_i} \in \mathbf{\Sigma}_{j_i}$. The union V of two signals V_1, V_2 takes the set union, i.e., $V = V_1 \cup V_2$. We will slightly abuse notation V to represent the random variable of a signal.

• *Data-generating process*. A data-generating process is a joint distribution $\pi \in \Delta(\mathbf{\Sigma}_1 \times \ldots \times \mathbf{\Sigma}_n \times \mathbf{\Omega})$ over the basic signals and the payoff-relevant state. However, the DM may only observe a subset V of the n basic signals. Conditioning on receiving a signal $V = v$, the DMs who know the data-generating process is able to infer the Bayesian posterior $\Pr[\omega|v]$ of the state, thus improving their payoff. Slightly abusing the notation, we will write $\pi(v, \omega)$ as the marginal probability of the signal realized to be v and the state being ω with expectation over unobserved signals.

2.1 Rational Decision Maker

We suppose a rational DM who knows the data-generating process, observes a signal realization, updates their prior to arrive at posterior beliefs, and then chooses a decision to maximize their expected payoff based on the posterior belief. Formally, the rational DM's expected payoff given a (set of) signals V is

$$\mathrm{R}(V) = \mathbf{E}_{v\sim\pi}[\max_{d\in\mathbf{D}} \mathbf{E}_{\omega\sim\Pr(\omega|v)}[S(d,\omega)]]$$

We use $\emptyset$ to represent a null signal, such that $\mathrm{R}(\emptyset)$ is the expected payoff of a Bayesian rational DM who has no access to a signal but only uses their prior belief to make decisions. In this case, the Bayesian rational DM will take the best fixed action and their expected payoff is

$$\mathrm{R}(\emptyset) = \max_{d\in\mathbf{D}} \mathbf{E}_{\omega\sim\pi}[S(d,\omega)]$$

Given a set of signals V_1 and a ground set of signals V_2, we can define the *information gain* from V_1 over V_2, the payoff improvement of V_1 over the payoff obtainable from V_2.

$$\gamma(V_1; V_2) = \mathrm{R}(V_1 \cup V_2) - \mathrm{R}(V_2). \tag{1}$$

2.2 Information in Behavioral Decisions

We use the term "behavioral DM" for a human who makes the decision in a decision-making problem after observing the signals. The intuition behind our approach is that any information that is used by behavioral DMs should eventually reveal itself through variation in their behaviors. Therefore, the information value in behavioral decisions can be recovered by offering the behavioral decisions as a signal to the Bayesian rational DM, which is equivalent to the information gain from behavioral decisions over a null signal. Similarly, we can look to the information gain from different signals over the behavioral decisions alone to test how useful the signals are beyond the information revealed in the behavioral decisions.

We model the decisions of a behavioral DM as a random variable D^b from the action space $\mathbf{D}$, which follows the distribution $\pi^b \in \Delta(\mathbf{\Omega} \times \mathbf{\Sigma}_1 \times \ldots \times \mathbf{\Sigma}_n \times \mathbf{D})$ – the joint behavior of the human correlated with the state and signals. The Bayesian rational DM knows the joint distribution π^b. After observing human decisions, the rational DM updates to a posterior and selects the decision that maximizes their expected payoff. Their expected payoff is given by the function:

$$\mathrm{R}(D^b) = \mathbf{E}_{d^b\sim\pi^b}[\max_{d\in\mathbf{D}} \mathbf{E}_{\omega\sim\Pr(\omega|D^b=d^b)}[S(d,\omega)]]$$

Information Gain of Signals Over Behavioral Decisions We seek to identify signals that can potentially improve behavioral decisions by analyzing their expected information gain $\gamma(V; D^b)$, the improvement in payoff expected from having the signal V over only having the behavioral action D^b. If the information gain of a signal is low over having only the behavioral decisions, this means either that the behavioral DM has already exploited the information, or that the information value to the decision problem of the signal is low. If, however, the information gain of a signal is high, then in theory the behavioral DM can improve their payoff by incorporating the signal's information in their decision making.

However, the information value of a basic signal may be overlooked if its value in combination with other signals is not considered. Signals can be complemented [Chen and Waggoner, 2016], i.e, they contain no information value by themselves but a considerable value when combined with other signals. For example, two signals Σ_1 and Σ_2 ight be uniformly random bits and the state $\omega = \Sigma_1 \oplus \Sigma_2$, the XOR of Σ_1 and Σ_2. In this case, neither of the signals offers information value on its own but knowing both can lead to the maximum payoff. To consider this complementation between signals, we use the Shapley value ϕ [Shapley, 1953] to interpret the contribution to information gain of each basic signal. The Shapley value calculates the average of the marginal contribution of a basic signal $\mathbf{\Sigma}_i$ in every combination of signals.

$$\phi(\Sigma_i) = \frac{1}{n} \sum_{V \subseteq \{V_1,\ldots,V_n\}/\{\Sigma_i\}} \binom{(n-1)}{|V|}^{-1} (\gamma(V \cup \{\Sigma_i\}; D^b) - \gamma(V; D^b)) \tag{2}$$

The Shapley value suggests how much information value of the basic signal is unexploited by the behavioral DM on average in all combinations.

Information Gain of Behavioral Decisions Over Signals We analyze the additional information contained in behavioral decisions beyond what is contained in the other available signals by examining the information gain of the behavioral decision over the signal, denoted as $\gamma(D^b; V)$. Suppose the set of all signals formalized in the data-generating process are $\overline{V} = \{\Sigma_1, \ldots, \Sigma_n\}$. $\gamma(D^b; \overline{V})$ captures the value of information that is reflected in behavioral decisions beyond the signals formalized by the data-generating process.

Our framework also offers a way to assess whether humans bring addition relevant information over an AI model for a decision task. Denote the AI predictions and human behavioral decisions as random variables D^{AI} and D^H. $\gamma(D^H; D^{AI})$ gives the value of additional information that is reflected in human decisions beyond AI predictions for the decision task.

3 Experiment

We apply our model to a deepfake video detection task studied by Groh et al. [2022], where participants are asked to judge whether a video is genuine or has been manipulated by neural network models. They are given access to predictions from a computer vision model that achieved an accuracy score of 65% on 4,000 videos in heldout data. Participants first review the video and report an initial decision. Then, in a second round, they are told the AI's recommendation and choose whether to change their initial decision. Participants are asked to report their belief that the video is fake in 1% increments: $d \in \{0\%, 1\%, \ldots, 100\%\}$.

We use the Brier score as the payoff function in our model: $S(\omega, d) = 1 - (\omega - d)^2$, with the binary payoff-related state: $\omega \in \{0, 1\} = \{\text{genuine}, \text{fake}\}$. We construct the basic signals in our model by the seven video-level features proposed by Groh et al. [2022]: graininess, blurriness, darkness, presence of a flickering face, presence of two people, presence of a floating distraction, and the presence of an individual with dark skin, all of which are hand-labeled as binary indictors. We estimate the data-generating process using the realizations of signals, state and behavioral decisions in the experiment data of Groh et al. [2022].

We show the results in Figure 1, where each distribution shows the distribution of the information gain of the signal over behavioral decisions. The signals are on the y axis and behavioral decisions are encoded by different colors. The information gain is on the payoff scale, which is bounded by $[0, 1]$, where 1 means the signal can improve the payoff of a rational decision maker who performs as badly as possible (defined by the scoring rule, e.g., 0 payoff in Brier score) to a rational decision maker who achieves the maximum payoff (e.g., 1 payoff in Brier score).

Unaided human v.s. AI. We first compare how participants without AI assistance and the AI use information in the deepfake detection task. Specifically, we calculate the Shapley value of information gain, $\phi^{D^H}(i)$ for participants and $\phi^{D^{AI}}(i)$ for AI, as shown in Equation 2, for each video-level feature V_i. The information gain over behavioral decisions reflects the information value of signals that are not fully redundant with the information in the behavioral decisions.

First, we observe that, in general, the information gain of the features over the AI decisions are lower than those for human decisions. There are several exceptions, such as the presence of an individual

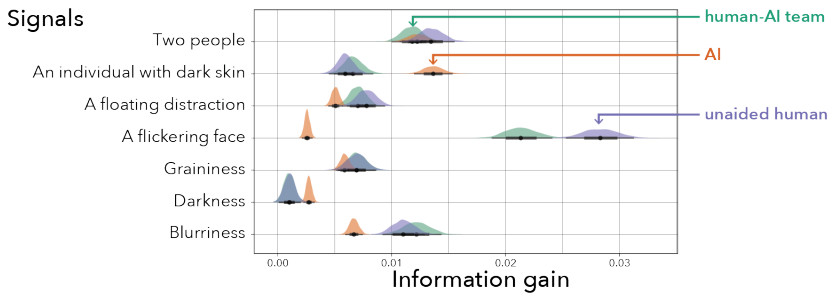

Figure 1: Information gain over the decisions of human-AI team, AI and unaided human.

with dark skin. This suggests that, overall, the information in features is better exploited by AI than by participants. More specifically, we find that the AI uses certain features much more effectively than participants. For instance, the presence of a flickering face offers the least information gain over AI decisions among all the features, whereas it is the feature that offers the largest information gain over human decisions. This suggests that one way to improve the current human-AI performance is to help the participants better exploit the information that AI exploits well but participants did not. Second, we find that the AI relies on less sensitive information compared to participants. For example, AI uses the presence of an individual with dark skin the least among all features, while for participants it is the second most important feature.

Unaided human v.s. human-AI team. We assess the information gain after participants are presented with AI recommendations in the deepfake detection task. We calculate the Shapley value of information gain $\phi^{D^{HAI}}(i)$ for human participants with AI recommendations relative to without. We find that simply displaying the AI's predictions to participants does not necessarily help them better exploit the potential value of information that they exploited poorly without access to the AI. For example, even though the information gain of the presence of a flickering face is reduced when presenting participants with AI predictions relative to without, the AI's much smaller gain for the signal implies participants could still use it much more effectively. This suggests that better interventions (e.g., explanations for AI predictions) may be needed to help people better incorporate some signals. Second, for the signals that the AI does not exploit well, offering the AI predictions does not necessarily reduce participants' usage of that information. For example, for the signal denoting the presence of an individual with dark skin, we did not see a significant improvement on the information gain over human-AI decisions compared to the gain over human decisions. Both of these findings suggest that simply displaying AI predictions may not change people's usage of information and improve their decision quality. Other interventions (such as explanations) should be explored to improve the use of potentially valuable by humans in human-AI collaborations.

4 Discussion

In this paper, we propose an approach to investigating information value in the context of a human and AI paired on a decision task. We identify the unexploited value of information contained in available signals by quantifying the information gain (as the increase in the expected payoff of a rational decision maker) of the signals over the behavioral decisions. We also identify the information value reflected in the behavioral decisions that is not contained in the signals. This approach makes it possible to identify signals that could be used to improve behavioral decisions through further interventions. For example, explanations might be designed to focus attention on signals whose information gain is low over the AI predictions but high over the human decisions.

Our work has limitations. First, the problem of complementation between signals is still not fully addressed. Though we develop a model using Shapley value to calculate the marginal contribution of signals in all combinations, we only account for the complementation and substitution between observed signals. Second, our methodology investigates information value in terms of the improvement that an ideal decision maker can achieve given that information.

Acknowledgments and Disclosure of Funding

We thank Groh et al. [2022], who provided their data for demonstration in this paper. We thank the annonymous reviewers for feedback.

References

Stefanía Ægisdóttir, Michael J White, Paul M Spengler, Alan S Maugherman, Linda A Anderson, Robert S Cook, Cassandra N Nichols, Georgios K Lampropoulos, Blain S Walker, Genna Cohen, et al. The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, 34(3):341–382, 2006.

Rohan Alur, Manish Raghavan, and Devavrat Shah. Distinguishing the indistinguishable: Human expertise in algorithmic prediction. *arXiv preprint arXiv:2402.00793*, 2024.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445717. URL `https://doi.org/10.1145/3411764.3445717`.

David Blackwell et al. Comparison of experiments. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, volume 1, page 26, 1951.

Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable ai systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, IUI '20, page 454–464, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371186. doi: 10.1145/3377325.3377498. URL `https://doi.org/10.1145/3377325.3377498`.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction*, 5(CSCW1):1–21, 2021.

Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*, pages 160–169, Oct 2015. doi: 10.1109/ICHI.2015.26.

Yiling Chen and Bo Waggoner. Informational substitutes. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 239–247. IEEE, 2016.

Teppo Felin and Matthias Holweg. Theory is all you need: Ai, human cognition, and causal reasoning.

Riccardo Fogliato, Alexandra Chouldechova, and Zachary Lipton. The impact of algorithmic risk assessments on human predictions and its analysis via crowdsourcing studies. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–24, 2021.

Raymond Fok and Daniel S Weld. In search of verifiability: Explanations rarely enable complementary performance in ai-advised decision making. *AI Magazine*, 2023.

Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, 2019.

Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, 119(1):e2110013119, 2022.

William M Grove, David H Zald, Boyd S Lebow, Beth E Snitz, and Chad Nelson. Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, 12(1):19, 2000.

Ziyang Guo, Yifan Wu, Jason D Hartline, and Jessica Hullman. A decision theoretic framework for measuring ai reliance. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 221–236, 2024.

Jessica Hullman, Alex Kale, and Jason Hartline. Decision theoretic foundations for experiments evaluating human decisions. *arXiv preprint arXiv:2401.15106*, 2024.

Maia Jacobs, Melanie F Pradier, Thomas H McCoy Jr, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry*, 11(1):108, 2021.

Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1):89–109, 2001.

Vivian Lai and Chenhao Tan. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 29–38, 2019.

Paul E Meehl. *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. University of Minnesota Press, 1954.

Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2, 1953.

Michelle Vaccaro and Jim Waldo. The effects of mixing machine learning and human judgment. *Communications of the ACM*, 62(11):104–110, 2019.

Yifan Wu, Ziyang Guo, Michalis Mamakos, Jason Hartline, and Jessica Hullman. The rational agent benchmark for data visualization. *IEEE transactions on visualization and computer graphics*, 2023.