Proceedings of the Eleventh AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2023)

A Taxonomy of Human and ML Strengths in Decision-Making to Investigate Human-ML Complementarity

Charvi Rastogi*, Liu Leqi*, Kenneth Holstein, Hoda Heidari

School of Computer Science, Carnegie Mellon University
crastogi, leqil, kjholste, hheidari @cs.cmu.edu

Abstract

Hybrid human-ML systems increasingly make consequential decisions in a wide range of domains. These systems are often introduced with the expectation that the combined human-ML system will achieve *complementary performance*, that is, the combined decision-making system will be an improvement compared with either decision-making agent in isolation. However, empirical results have been mixed, and existing research rarely articulates the sources and mechanisms by which complementary performance is expected to arise. Our goal in this work is to provide conceptual tools to advance the way researchers reason and communicate about human-ML complementarity. Drawing upon prior literature in human psychology, machine learning, and human-computer interaction, we propose a taxonomy characterizing distinct ways in which human and ML-based decision-making can differ. In doing so, we conceptually map potential mechanisms by which combining human and ML decision-making may yield complementary performance, developing a language for the research community to reason about design of hybrid systems in any decision-making domain. To illustrate how our taxonomy can be used to investigate complementarity, we provide a mathematical aggregation framework to examine enabling conditions for complementarity. Through synthetic simulations, we demonstrate how this framework can be used to explore specific aspects of our taxonomy and shed light on the optimal mechanisms for combining human-ML judgments.

1 Introduction

In recent years, we have witnessed a rapid growth in the deployment of machine learning (ML) models in decision-making systems across a wide range of domains, including healthcare (Patel et al. 2019; Rajpurkar et al. 2020; Tschandl et al. 2020; Bien et al. 2018), credit lending (Bussmann et al. 2021; Kruppa et al. 2013), criminal justice (Angwin et al. 2016; Kleinberg et al. 2018), and employment (Raghavan et al. 2020; Hoffman, Kahn, and Li 2017). For example, in the criminal justice system, algorithmic recidivism risk scores inform pre-trial bail decisions for defendants (Angwin et al. 2016). In credit lending, lenders routinely use credit-scoring models to assess the risk of default by applicants (Kruppa et al. 2013). The excitement around modern ML systems facilitating high-stakes decisions is fueled by the promise of these technologies to tap into large datasets, mine the relevant statistical patterns within them, and utilize those patterns to make more accurate predictions at a lower cost and without suffering from the same cognitive biases and limitations as human decision-makers. Growing evidence, however, suggests that ML models are vulnerable to various biases (Angwin et al. 2016) and instability (Finlayson et al. 2018). Furthermore, they often produce harmful outcomes in practice, given that they lack humans strengths such as commonsense reasoning abilities, cognitive flexibility, and social and contextual knowledge (Alkhatib 2021; Holstein and Aleven 2021; Lake et al. 2017; Miller 2019). These observations have led to calls for both human and ML involvement in high-stakes decision-making systems—with the hope of combining and amplifying the respective strengths of human thinking and ML models through carefully designed *hybrid* decision-making systems. Such systems are common in practice, including in the domains mentioned above.

Researchers have proposed and tested various hybrid human-ML designs, ranging from human-in-the-loop (Russakovsky, Li, and Li 2015) to algorithm-in-the-loop (De-Arteaga, Fogliato, and Chouldechova 2020; Saxena et al. 2020; Brown et al. 2019; Green and Chen 2019) arrangements. However, empirical findings regarding the success and effectiveness of these proposals are mixed (Holstein and Aleven 2021; Lai et al. 2021). Simultaneously, a growing body of theoretical work has attempted to conceptualize and formalize these hybrid designs (Gao et al. 2021; Bordt and von Luxburg 2020) and study optimal ways of aggregating human and ML judgments within them (Madras, Pitassi, and Zemel 2018; Mozannar and Sontag 2020; Wilder, Horvitz, and Kamar 2020; Keswani, Lease, and Kenthapadi 2021; Raghu et al. 2019; Okati, De, and Gomez-Rodriguez 2021; Donahue, Chouldechova, and Kenthapadi 2022; Steyvers et al. 2022).

Much prior work has studied settings where the ML model outperforms the human decision-maker. These studies are frequently focused on tasks where there are no reasons to expect upfront that the human and the ML model will have complementary strengths (Bansal et al. 2021; Guerdan et al. 2023; Holstein and Aleven 2021; Lurie and Mulligan 2020). For example, some experimental studies employ

*Both authors contributed equally to this work.

untrained crowdworkers on tasks that require extensive domain expertise, without which there is no reason to expect that novices would have complementary strengths (Fogliato, Chouldechova, and Lipton 2021; Lurie and Mulligan 2020; Rastogi et al. 2022). Other experimental studies are designed in ways that artificially constrain human performance—for instance, by eliminating the possibility that humans and ML systems have access to complementary information (Guerdan et al. 2023). Meanwhile studies on human-ML decision-making in real-world settings such as healthcare (Tschandl et al. 2020; Patel et al. 2019) sometimes demonstrate better human-ML team performance than either agent alone. However, the *reasons* for complementary team performance are often left unexplained, where we define human-ML *complementarity* as the condition in which a combination of human and ML decision-making outperforms[1] both human- and ML-based decision-making in isolation.

We argue, therefore, that there is a clear need to form a deeper, more fine-grained understanding of what types of human-ML systems exhibit complementarity in combined decision-making. To respond to this gap in the literature, we build a novel *taxonomy* of relative strengths and weaknesses of humans and ML models in decision-making, presented in Figure 1. This taxonomy aims to provide a shared understanding of the causes and conditions of complementarity so that researchers and practitioners can design more effective hybrid systems and focus empirical evaluations on promising designs—by investigating and enumerating the distinguishing characteristics of human vs. ML decision-making upfront. Our taxonomy covers application domains wherein the decision at stake is solely based on *predicting* some outcome of interest (Mitchell et al. 2018). Henceforth, we use the terms 'prediction' and 'decision' interchangeably. Some examples of predictive decisions are diagnosis of diabetic retinopathy (Gulshan et al. 2016), predicting recidivism for pretrial decisions (Dressel and Farid 2018), and consumer credit risk prediction (Bussmann et al. 2021).

To build our taxonomy of human-ML complementarity, we surveyed the literature on human behavior, cognitive and behavioral sciences, as well as psychology to understand the essential factors across which human and ML decision-making processes differ. Following traditions in cognitive science and computational social science (Lake et al. 2017; Marr and Poggio 1977), we understand human and ML decision-making through a computational lens. Our taxonomy maps distinct ways in which human and ML decision-making can differ (Section 3).

To illustrate how our taxonomy can be used to investigate when we can expect complementarity in a given setting and what modes of human-ML combination will help achieve it, we present a mathematical framework that captures each factor in the taxonomy. In particular, we formalize an optimization problem for convex combination of human and ML decisions. This problem setup establishes a pathway to help researchers explore which characteristics of humans and ML models have the potential to foster complementary performance. To categorize different types of complementarity, we propose quantitative measures of complementarity, designed to capture two salient modes of human-ML collaboration in the literature: routing (or deferral) and communication-based collaboration. To demonstrate the use of our taxonomy, the optimization problem setup, and the associated metrics of complementarity, we simulate optimal human-ML combinations under two distinct conditions: (1) human and ML models have access to different feature sets, (2) human and ML models have different objective functions. By comparing optimal aggregation strategies under these conditions, we gain critical insights regarding the contribution of each decision-making agent towards the optimal combined decision. This informs the effective design of human-ML partnerships under these settings for future research and practice. Taken together, this work highlights that combining human-ML judgments should leverage the unique strengths and weaknesses of each entity, as different sources of complementarity impact the extent and nature of performance improvement achievable through human-ML collaboration.

In summary, this paper contributes a unifying taxonomy and formalization for human-ML complementarity. Our taxonomy characterizes major differences between human and ML predictions, and our optimization-based framework formally characterizes optimal aggregation of human and machine decisions under various conditions and the type of complementarity that produces optimal decisions. With these contributions, we hope to provide a common language and an organizational structure to inform future research in this increasingly important space for human-ML combined decision-making.

2 Methodology for Designing the Taxonomy

To investigate the potential for complementarity in human-ML combined decision-making, we need to understand the respective strengths and drawbacks of the human decision-maker and the ML model in the context of the application. For instance, it has been observed that while ML models draw inferences based on much larger bodies of data than humans could efficiently process (Jarrahi 2018), human decision-makers bring rich contextual knowledge and common sense reasoning capabilities (Holstein and Aleven 2021; Miller 2019; Lake et al. 2017) to the decision-making process, which ML models may be unable to replicate. Thus, we develop a taxonomy for human-ML decision-making that accounts for broad differences between human decision-makers and machine learning, encompassing applications with predictive decision-making.

To inform this taxonomy, we draw from existing syntheses in human psychology, machine learning, AI, and human-computer interaction to understand distinguishing characteristics that have been observed between human decision-makers and ML in the context of predictive decision-making. In cognitive science, Lake et al. (2017) review major gaps between human and ML capabilities by synthesizing existing scientific knowledge about human abilities that have thus far defied automation. In management science, Shrestha, Ben-Menahem, and Von Krogh (2019)

[1]Complementary performance may present along any performance metric, and does not necessarily refer to accuracy.

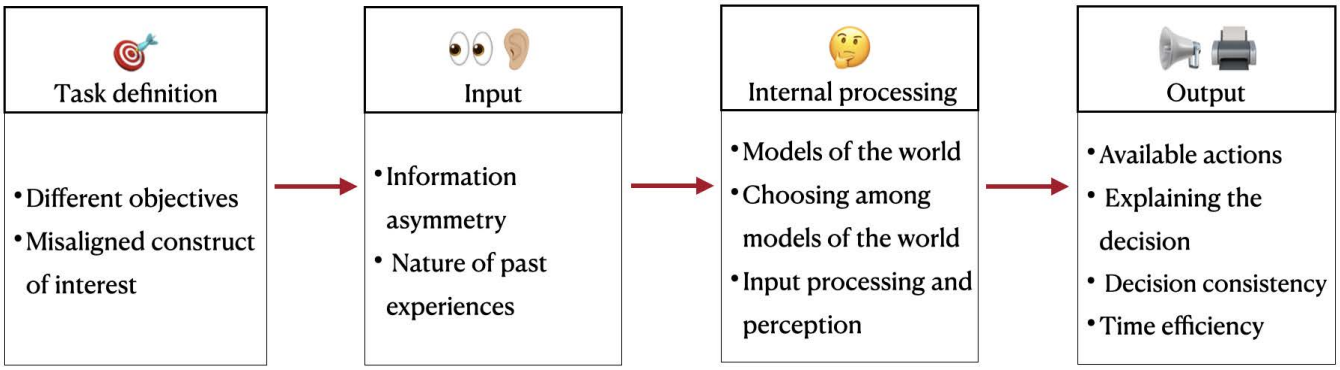

Figure 1: Proposed taxonomy of human and ML strengths & weaknesses in decision-making.

identify characteristics of human and ML decision-making along four key axes: decision space, process and outcome interpretability, speed, and replicability, and discuss their combination for organizational decision-making. In human-computer interaction, Holstein, Aleven, and Rummel (2020) conceptually map distinct ways in which humans and AI can augment each others' abilities in real-world teaching and learning contexts. More recently, Lai et al. (2021) surveyed empirical studies on human-AI decision-making to document trends in study design choices (e.g., decision tasks and evaluation metrics) and empirical findings. We draw upon this prior literature to summarize key differences in human and ML-based predictive decision-making across multiple domains, with an eye towards understanding opportunities to combine their strengths.

Computational lens. Our taxonomy takes a computational perspective towards analysing human and ML decision-making. As with any modeling approach or analytic lens, computational-level explanations are inherently reductive, yet are often useful in making sense of complex phenomena for this very reason. Computational-level explanations often provide an account of a *task* an agent performs, the *inputs* that the agent takes in, the ways in which the agent *perceives and processes* these inputs, and the kinds of *outputs* that the agent produces. Accordingly, our taxonomy is organized into four elements: (1) task definition, (2) input, (3) internal processing, and (4) output.

We now provide mathematical notation to clearly express the computational perspective of decision-making in our taxonomy. Formally, the agent's decision-making setting is specified by a feature space $\mathcal{X}$, an action space $\mathcal{A}$, and the space of observed outcomes, $\mathcal{O}$. At a high-level, the agent perceives an instance $\mathbf{X} \in \mathcal{X}$, chooses to take an action $a \in \mathcal{A}$ based on its relevant prior knowledge and experiences, and observes an outcome $O \in \mathcal{O}$ as a result. To emphasize that the outcome O is influenced by $\mathbf{X}$ and a, we slightly abuse the notation to denote the outcome of action a on instance $\mathbf{X}$ as $O(\mathbf{X}, a)$. We consider the agent's perception of an instance $\mathbf{X}$ to be denoted by $s(\mathbf{X})$, where $s : \mathcal{X} \to \mathcal{X}$. Next, the agent's prior knowledge and relevant experiences are assumed to be encompassed in a set $\mathcal{D}$. The goal of the decision-making agent is to choose a policy, $\pi : \mathcal{X} \to \mathcal{A}$, in the space of feasible policies Π, such that π leads to favorable overall outcome quality, measured by an evaluation function F. Here F takes in a policy and outputs a real number. For instance, the expected outcome of a policy is a common choice for F. Finally, the agent chooses their optimal policy $\overline{\pi}$ using their optimization process OPT for choosing among feasible policies that lead to favorable F values. Using the categorization and mathematical formalization above, and drawing upon relevant background literature as presented in this section, we now provide our taxonomy for relative human and ML strengths.

3 A Taxonomy of Human and ML Strengths & Weaknesses in Decision-Making

In this work, we consider two decision-making agents, the human and the ML model denoted respectively by H and M. Building upon the notation in Section 2, we denote the feature space available to each agent by $\mathcal{X}_{\text{H}}, \mathcal{X}_{\text{M}}$ correspondingly, where $\mathcal{X}_{\text{H}}, \mathcal{X}_{\text{M}} \subseteq \mathcal{X}$. Similarly, for each variable introduced for our decision-making setting in the previous section, we consider a human version and a ML version, denoted by subscript H and M respectively. We now present our taxonomy, visually represented in Figure 1.

3.1 Task Definition

We now describe the distinguishing characteristics that have been observed in the definition of the decision-making task used by the human and the ML model.

- **Objective.** Most machine learning models aim to only optimize the expected performance, e.g., minimize the expected loss for supervised learning models and maximize the expected cumulative rewards for reinforcement learning models. While recent research has explored ways to build models with respect to a more diverse set of objectives, including different risk measures (Leqi, Prasad, and Ravikumar 2019a; Khim et al. 2020), fairness definitions (Chouldechova and Roth 2020) and interpretability notions (Lipton 2018; Miller 2019), it is commonly difficult or impractical to encode all aspects of the objectives that a human decision-maker would aim to optimize (Kleinberg et al. 2018). Using our notation, this is expressed as $F_{\text{H}} \neq F_{\text{M}}$. For example, when making a lending

decision, in addition to considering various risk factors, bankers may also care about aspects such as maintaining their relationships with clients and specific lending practices in their organization (Trönnberg and Hemlin 2014).

- **Misaligned construct of interest.** ML models deployed in social contexts often involve theoretical constructs that are not directly observable in the data, such as socioeconomic status, teacher effectiveness, and risk of recidivism, which cannot be measured directly. Instead they are inferred indirectly via proxies: measurements of properties that are observed in the data available to a model. The process of defining proxy variables for a construct of interest necessarily involves making simplifying assumptions, and there is often a considerable conceptual distance between ML proxies and the ways human decision-makers think about the targeted construct (Green and Chen 2021; Guerdan et al. 2023; Jacobs and Wallach 2021; Kawakami et al. 2022). In other words, $O_{\mathrm{H}}(\mathbf{X}, a) \neq O_{\mathrm{M}}(\mathbf{X}, a)$. Jacobs and Wallach (2021) argue that several harms studied in the literature on fairness of sociotechnical systems are direct results of the mismatch between the construct of interests and the inferred measurements. For example, Obermeyer et al. (2019) examined racial biases in an ML-based tool used in hospitals. They found that the use of an indirect proxy (healthcare costs incurred by a patient) to predict patients' need for healthcare contributed to worse healthcare provision decisions for black versus white patients. In this example, although the proxy used (the monetary cost of care) was conveniently captured in available data, it differs significantly from the way healthcare professionals conceptualize patients' actual need for care.

3.2 Input

We now describe the distinguishing characteristics observed in the inputs used by humans and ML models.

- **Access to different information.** From the input perspective, in many settings such as healthcare, criminal justice, humans and machines have access to both shared and non-overlapping information: $\mathcal{X}_{\mathrm{H}} \neq \mathcal{X}_{\mathrm{M}}$. This is because real-world decision-making contexts often contain features of importance that cannot be codified for ML. For example, a doctor can see the physical presentation of a patient and understand their symptoms better, since this information is hard to codify and provide to the machine. Similarly, a judge learns about the predisposition of the defendant through interaction (Kleinberg et al. 2018). This phenomena is also referred to as unobservables (Holstein et al. 2023) and information asymmetry (Hemmer et al. 2022) in the literature on human-ML complementarity.
- **Nature of past experiences.** The nature of embodied human experience over the course of a lifetime differs substantially from the training datasets used by modern ML systems: $\mathcal{D}_{\mathrm{H}} \neq \mathcal{D}_{\mathrm{M}}$. For example, ML models are often trained using a large number of prior instances of a specific decision-making task, but for each instance, the training data contains a fixed and limited set of information. This often does not reflect the richness of human experience. Humans make their decisions with reference to a lifetime of experiences across a range of domains, and it is difficult to explicitly specify the information they take into account. By contrast, ML models may learn from training data that comprise narrow slices from a vast number of human decision-makers' decisions, whereas humans typically learn only from their own experiences or from a small handful of other decision-makers.

3.3 Internal Processing

We now describe the distinguishing characteristics of the internal processes used by humans and ML systems.

- **Models of the world.** As is comprehensively overviewed in Lake et al. (2017), humans rely upon rich mental models and "theories" that encode complex beliefs about causal mechanisms in the world, not just statistical relationships. This results in humans having a different set of models of the world than those embodied by ML models: $\Pi_{\mathrm{H}} \neq \Pi_{\mathrm{M}}$. For example, starting from an early age, humans develop sophisticated systems of beliefs about the physical and social worlds (intuitive physics and intuitive psychology), which strongly guide how they perceive and make decisions in the world. In contrast to modern ML systems, humans' mental models tend to be compositional and causal. In turn, these strong prior beliefs about the world can enable humans to learn rapidly in comparison to modern ML systems, and to make inferential leaps based on very limited data (e.g., one-shot and few-shot learning) (Gopnik and Wellman 2012; Lake et al. 2017; Tenenbaum et al. 2011). On the other hand, the model class of the machine decision-maker has a more mathematically tractable form—whether it is a class of parametric or non-parametric models (Friedman 2017). Although when designing these models such as neural networks, researchers commonly encode domain knowledge through the data and the model architecture, most machine learning models still suffer from distribution shift (Quiñonero-Candela et al. 2009) and lack of interpretability (Gilpin et al. 2018), and require large sample sizes.
- **Input processing and perception.** The ways decision-makers perceive inputs is informed by their models of the world (Gentner and Stevens 2014; Holstein, Aleven, and Rummel 2020). Following research in human cognition and ML, we highlight three sources of variation in input perception: (1) differences in mental/computational capacity, (2) differences in human versus machine biases, and (3) tendencies towards causal versus statistical perception. Here the first implies $s_{\mathrm{H}} \neq s_{\mathrm{M}}$ and the remaining two indicate $\pi_{\mathrm{H}} \neq \pi_{\mathrm{M}}$. For instance, compared with ML systems, humans demonstrate less capacity to perceive small differences in numerical values (Amitay et al. 2013; Findling and Wyart 2021). Furthermore, both humans and ML systems can bring in both adaptive and maladaptive biases, based on their experiences and models of the world, which in turn shape the ways they process and perceive new situations (Fitzgerald and Hurst 2017; Wistrich and Rachlinski 2017; Kleinberg et al. 2018; Gentner and Stevens 2014). However, in some cases humans and ML systems may have complementary biases, opening room

for each to help mitigate or compensate for the other's limitations (Holstein, Aleven, and Rummel 2020; Tan et al. 2018). Finally, research on human cognition demonstrates that humans are predisposed to perceiving causal connections in the world, and drawing causal inferences based on their observations and interactions in the world (Gopnik and Wellman 2012; Lake et al. 2017). While these abilities can sometimes be understood by analogy to the kinds of statistical learning that most modern ML systems are based upon (Tenenbaum et al. 2011), other aspects of human causal cognition appear to be fundamentally different in nature (Lake et al. 2017). As with bias, these abilities can be a double-edged sword. In some scenarios, human causal perception may lead to faulty inferences based on limited data. By contrast, ML systems will sometimes have an advantage in drawing more reliable inferences based on statistical patterns in large datasets. In other settings, human causal perception can help to overcome limitations of ML systems. For example, in many instances, human decision-makers have been observed to be better than ML systems at adapting to out-of-distribution instances, through the identification and selection of causal features for decision-making (Lake et al. 2017).

- **Choosing among models of the world.** Given the task definition, models of the world, and data, ML models differ from humans in searching for the model that optimizes their objective: $\text{OPT}_\text{H} \neq \text{OPT}_\text{M}$. Modern ML models (e.g., neural networks) are commonly learned using first-order methods and may require a huge amount of computational resource due to the size of the models (Bottou 2010). On the other hand, humans may employ heuristics that can be executed in a relatively short amount of time (Simon 1979). These simple strategies may have advantages over more complex models when the inherent uncertainty in the task is high. For a more comprehensive review on when and how such heuristics may be more preferable, we refer readers to Kozyreva and Hertwig (2021).

3.4 Output

We now describe the distinguishing characteristics of the outputs generated by humans and ML systems.

- **Available actions.** In real-world deployment settings, the set of possible decisions or actions available to ML models versus humans can be different: $\mathcal{A}_\text{H} \neq \mathcal{A}_\text{M}$. For example, in the context of K-12 education, ML-based tutoring software may be able to provide just-in-time hints to students, to help struggling students them with math content. Meanwhile, although a human teacher working alongside this software in the classroom has limited time to spend with each student, they can take a wider range of actions to support students, such as providing emotional support or helping students with prerequisite content that lies outside of the software's instructional repertoire (Holstein, Aleven, and Rummel 2020). Similarly, in the context of ML-assisted child maltreatment screening, a model may only be able to recommend that a case be investigated or not investigated, based on the information that is currently available. By contrast, Kawakami et al. (2022) report that human call screeners may take actions to gather additional information as needed, e.g. by making phone calls to other stakeholders relevant to a case.
- **Explaining the decision.** Humans and ML have differing abilities in communicating the reasoning behind their decisions. There has been extensive research in explainability (XAI) and interpretability for ML (Adadi and Berrada 2018). Research in cognitive and social psychology observes that humans are generally better than ML algorithms at generating coherent explanations that are meaningful to other humans. Furthermore, Miller (2019) argues that XAI research should move away from imprecise, subjective notions of "good" explanations and instead focus on reasons and thought processes that people apply for explanation selection. They find that human explanations are contrastive, selected in a biased manner, and most importantly they are social and contextual. On the other hand, humans' explanations may not have a correspondence to their actual underlying decision processes (Nisbett and Wilson 1977), whereas with ML models we can always trace the precise computational steps that led to the output prediction (Hu, Rudin, and Seltzer 2019).
- **Uncertainty communication.** With increasing research in uncertainty quantification for machine learning, new methods have been devised for calibrating a ML model's uncertainty in its prediction (Abdar et al. 2021). Moreover, methods have been developed to decompose the model uncertainty into aleatoric uncertainty and epistemic uncertainty (Hüllermeier and Waegeman 2021), where aleatoric uncertainty signifies the inherent randomness in an application domain and cannot be reduced, and epistemic uncertainty, also known as systematic uncertainty, signifies the uncertainty due to lack of information or knowledge, and can be reduced. However, these uncertainty quantification methods may not necessarily be well-calibrated (Abdar et al. 2021), and are an active research direction. Meanwhile, human decision-makers also find it difficult to calibrate their uncertainty or their confidence in their decisions (Brenner, Griffin, and Koehler 2005), and tend to output discrete decisions instead of uncertainty scores. Moreover, different people have different scales for uncertainty calibration (Zhang and Maloney 2012).
- **Output consistency.** We define a given decision-maker to have a consistent output when they always produce the same output for the same input. Therefore, we consider the inconsistency in decisions that are based on factors independent of the input, we call them extraneous factors. Some examples of extraneous factors are the time of the day, the weather, etc. Research in human behavior and psychology has shown that human judgments show inconsistency (Kahneman et al. 2016). More specifically, there is a positive likelihood of change in outcome by a given human decision-maker given the exact same problem description at two different instances. Within-person inconsistency in human judgments has been observed across many domains, including medicine (Koran 1975; Kirwan et al. 1983), clinical psychology (Little 1961), finance and management (Kahneman et al. 2016). This form of incon-

sistency is not exhibited by standard ML algorithms.[2]

- **Time efficiency.** In many settings, ML models can generate larger volumes of decisions in less time than human decision-makers. In addition to potentially taking more time per decision, humans often have comparatively scarce time for decision-making overall.

4 Investigating the Potential for Human-ML Complementarity

To understand how the differences in human and machine decision-making result in complementary performance, we formulate an optimization problem to aggregate the human and the ML model outcomes. The key motivation here is to use information available about human and ML decision-making (in the form of historical data or decision-making models) to understand the potential for complementarity in human-ML joint performance. Specifically, this optimization problem outputs the optimal convex combination of the two decision-makers' outputs wherein the aggregation mechanism represents the best that the human-ML joint decision-making can achieve in our setting.

In our decision-making setting, as mentioned in Section 2, we consider a feature space $\mathcal{X}$, an action space $\mathcal{A}$ and an outcome space $\mathcal{O}$. Given a problem domain, the goal is to combine the two decision-makers policies to find a joint policy denoted by $\overline{\pi} : \mathcal{X} \to \mathcal{A}$ that maximizes the overall quality of the decisions based on evaluation function, F,

$$\overline{\pi} \in \arg\max_{\pi \in \Pi} \; F(\pi). \tag{1}$$

We note that the overall evaluation function F for the joint policy π may be different from that used by the human F_{H} or the ML model F_{M}. We assume the joint policy is obtained by combing human and machine policies π_{H} and π_{M} over n number of instances through an aggregation function. We consider the outcome space to be scalar $\mathcal{O} \subseteq \mathbb{R}$. Given $\pi_{\mathrm{H}} \in \Pi_{\mathrm{H}}$, $\pi_{\mathrm{M}} \in \Pi_{\mathrm{M}}$, for an instance $\mathbf{X}_i$ where $i \in [n]$, the joint policy $\pi \in \Pi$ is given by

$$\pi(\mathbf{X}_i) = w_{\mathrm{H}}^{(i)} \pi_{\mathrm{H}}(\mathbf{X}_i) + w_{\mathrm{M}}^{(i)} \pi_{\mathrm{M}}(\mathbf{X}_i), \tag{2}$$

for some weights $w_{\mathrm{H}}^{(i)}, w_{\mathrm{M}}^{(i)} \in [0, 1]$ and $w_{\mathrm{H}}^{(i)} + w_{\mathrm{M}}^{(i)} = 1$ for all $i \in [n]$. Here note that we assume that the joint decision $\pi(\mathbf{X}_i)$ is a convex combination of the individual decisions $\pi_{\mathrm{H}}(\mathbf{X}_i)$ and $\pi_{\mathrm{M}}(\mathbf{X}_i)$. This assumption arises naturally to ensure that the joint decision lies between the human's and machine's decision. For a decision-maker (say human), the weight assigned for instance i, $w_{\mathrm{H}}^{(i)}$ indicates the amount of contribution from them towards the final decision: when $w_{\mathrm{H}}^{(i)} = 0$, the joint decision does not follow human's decision at all on instance $\mathbf{X}_i$, while $w_{\mathrm{H}}^{(i)} = 1$ indicates that their decision is followed entirely. For the optimal policy $\overline{\pi}$ defined in (1), its corresponding optimal weights are denoted by $\overline{w}_{\mathrm{H}}^{(i)}$ and $\overline{w}_{\mathrm{M}}^{(i)}$.

[2]There exists the special case of randomized models, we consider these outside the scope of our work and, further note that these models can be directly mapped to deterministic models with decision-based thresholds.

Several existing works on human-ML combination for decision-making, such as Donahue, Chouldechova, and Kenthapadi (2022); Raghu et al. (2019); Mozannar and Sontag (2020); Gao et al. (2021) are subsumed by our convex combination optimization setup. Particularly, our aggregation mechanism captures two salient modes: (1) The mode where an instance is routed to either the human or the ML decision maker, also known as deferral. This is represented by $w_{\mathrm{H}}^{(i)}, w_{\mathrm{M}}^{(i)} \in \{0, 1\}$ for all $i \in [n]$. (2) The mode where a joint decision lying between the human and the ML decision is applied to each instance. This is represented by $w_{\mathrm{H}}^{(i)}, w_{\mathrm{M}}^{(i)} \in (0, 1)$ for all $i \in [n]$.

4.1 Metrics for Complementarity

The proposed aggregation framework is a way to inspect the extent of complementarity in human-ML joint decision-making. Recall that, based on our definition, The joint policy π defined in (2) exhibits complementarity if and only if

$$F(\pi) > \max\{F(\pi_{\mathrm{H}}), F(\pi_{\mathrm{M}})\}.$$

Although this criterion provides a binary judgment on whether complementarity exists in a particular joint decision-making setting, it cannot be used to compare the amount of potential for complementarity in different settings. For instance, between two settings where machine can improve the performance of the human decision-maker on one instance versus on all instances, one may say that there is more complementarity exhibited in the second setting. Further, it does not distinguish between the two salient modes of combination defined above, where the second mode may require more interaction between the human and the machine decision-maker. So, to investigate the potential for complementarity in different settings more thoroughly, we introduce metrics for quantifying the complementarity between the human and ML decision-maker.

Specifically, we introduce the notion of within- and across-instance complementarity to represent the two modes of combination where for an instance $\mathbf{X}_i$, we either have only one of human or ML contributing to the final decision ($w_{\mathrm{M}}^{(i)} = 1$ or $w_{\mathrm{H}}^{(i)} = 1$), or both decision-makers contributing to the final decision partially ($w_{\mathrm{M}}^{(i)} > 0$ and $w_{\mathrm{H}}^{(i)} > 0$). These two types of combinations represent two ways of achieving complementarity. In the first one, there is no complementarity within a single task instance, since only the human or the ML model decision gets used. In this scenario, if the human and ML model provide the final decision for different instances of the task, we call this *across-instance complementarity*. In the second one, if both human and ML model contribute to the same instance $\mathbf{X}_i$, we call this *within-instance complementarity*. These two metrics help distinguish between different instance allocation strategies in human-ML teams described in (Roth et al. 2019). Formally, given the weights assigned to the two agents in the final decision, we define the two metrics as follows:

- **Across-instance complementarity** quantifies the variability of the human (or the machine) decision-maker's contribution to the final decision across all task instances.

Therefore, we define it as the variance of the weights assigned, written as

$$c_{\text{across}}(w_{\text{M}}, w_{\text{H}}) \coloneqq \frac{1}{n}\sum_{i=1}^{n}\left(w_{\text{M}}^{(i)} - \frac{1}{n}\sum_{i=1}^{n} w_{\text{M}}^{(i)}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(w_{\text{H}}^{(i)} - \frac{1}{n}\sum_{i=1}^{n} w_{\text{H}}^{(i)}\right)^2. \quad (3)$$

The equality follows directly using the constraint $w_{\text{M}}^{(i)} + w_{\text{H}}^{(i)} = 1$. In case of no variability across instances, that is if for both decision-makers, we have $w_{\text{M}}^{(i)}$ (or $w_{\text{H}}^{(i)}$) to be a constant for all $i \in [n]$, then $c_{\text{across}}(w_{\text{M}}, w_{\text{H}}) = 0$. The notion of across-instance complementarity is shown by works on decision deferral or routing including Mozannar and Sontag (2020); Madras, Pitassi, and Zemel (2018).

- **Within-instance complementarity** quantifies the extent of collaboration between the two decision-makers on each individual task instance. Formally, we define

$$c_{\text{within}}(w_{\text{M}}, w_{\text{H}}) \coloneqq 1 - \frac{1}{n}\sum_{i=1}^{n}\left(w_{\text{H}}^{(i)} - w_{\text{M}}^{(i)}\right)^2. \quad (4)$$

Importantly, the definition of within-instance complementarity satisfies some key properties: $c_{\text{within}}(w_{\text{H}}, w_{\text{M}})$ is maximized at $w_{\text{H}}^{(i)} = w_{\text{M}}^{(i)} = 0.5$ and minimized at $w_{\text{H}}^{(i)} \in \{0, 1\}$ for all $i \in [n]$. Thus, it is maximized when each decision-maker contributes equally and maximally to a problem instance and minimized when there is no contribution from one of the decision-makers. Further, it increases monotonically as $w_{\text{H}}^{(i)}$ and $w_{\text{M}}^{(i)}$ get closer to each other in value, that is the two decision-makers' contributions to the final decision get closer to half. This notion of complementarity is demonstrated in several works including Patel et al. (2019); Tschandl et al. (2020).

To have a better grasp on the above two metrics, and to understand the importance of each metric in measuring complementarity, we provide some demonstrative examples. Consider a simple setting with $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_4\}$ where each instance is equally likely, that is, $\mathbb{P}(\mathbf{X} = \mathbf{x}_i) = 1/4$ for all $i \in [4]$. The values of the two metrics under different aggregation weights are given below:

1. If $w_{\text{H}}^{(1)} = w_{\text{H}}^{(2)} = w_{\text{H}}^{(3)} = w_{\text{H}}^{(4)} = 0$, then $c_{\text{within}} = 0$, and $c_{\text{across}} = 0$.
2. If $w_{\text{H}}^{(1)} = w_{\text{H}}^{(2)} = 0, w_{\text{H}}^{(3)} = w_{\text{H}}^{(4)} = 1$, then $c_{\text{within}} = 0$, and $c_{\text{across}} = 0.25$.
3. If $w_{\text{H}}^{(1)} = w_{\text{H}}^{(2)} = w_{\text{H}}^{(3)} = w_{\text{H}}^{(4)} = 0.3$, then $c_{\text{within}} = 0.84$, and $c_{\text{across}} = 0$.

We note that although the second example has $c_{\text{within}} = 0$ and $c_{\text{across}} > 0$, which is the opposite of the third example, both the examples demonstrate complementarity. This shows that each metric introduced captures aspects of human-ML complementarity that is not captured by the other metric.

5 Synthetic Experiments to Illustrate Complementarity

In this section, we illustrate how our proposed framework can be used to investigate the extent and nature of complementarity via simulations. These simulations utilize human and ML models learned from data, where the two decision-makers have different access of information or they pursue different objectives. By quantifying the extent of different types of complementarity (i.e., within-instance and across-instance), we show how the proposed taxonomy and complementarity metrics can guide the research and practice of hypothesizing about and testing for complementarity with different types of human and ML decision-makers. To conduct these simulations, we choose specific aspects from our taxonomy in Section 3 and measure complementarity in the presence of corresponding differences between the human and the ML model. We note that these simulations are meant to be an illustrative and not exhaustive exploration of human-ML complementarity conditions that can be explored using the taxonomy.

Synthetic simulation setup. We consider a linear model for the data generating process: the features $\mathbf{X} \in \mathbb{R}^d$ are distributed as $\mathbf{X} \sim \mathcal{N}(0, \mathbf{I}_{d\times d})$; the target is given by $Y = \mathbf{X}^\top \beta + \epsilon$ where $\beta = (1 \cdots 1) \in \mathbb{R}^d$ and $\epsilon \sim \mathcal{N}(0, 1)$. For any given instance $\mathbf{X} \in \mathbb{R}^d$, both the ML and human decision-maker make a prediction using their respective linear model, which serves as a decision. We assume that the outcome for a given instance is determined by the squared loss incurred by the decision. For example, for the machine, given the true target Y and the prediction $\pi_{\text{M}}(\mathbf{X})$, the outcome is given by $O = (\pi_{\text{M}}(\mathbf{X}) - Y)^2$. The dimension of the features is chosen to be $d = 10$ for all simulations.

In Section 5.1, we study how human-ML complementarity varies when the human and the ML model have different feature information available to them; and in Section 5.2, the difference between the human and the machine arises via difference in objective functions for learning their respective policies. In the following simulations, we first use a training set of sample size $8,000$ to learn the respective optimal linear model for the human and the ML policy. Once the decision-makers' policies are learned, a separate testing set of size $2,000$ is used to compute and analyse the optimal aggregation weights. On this set, we measure and report the metrics of complementarity defined in Section 4.1.

5.1 Access to Different Feature Sets

First, we consider the setting where the human and the machine decision-maker have different information available to them. This is a potential source of complementarity in human-ML joint decision-making as mentioned in our taxonomy in Section 3 based on the input. To analyze the impact of information asymmetry on human-ML complementarity, we conduct synthetic experiments based on the general setup described at the beginning of Section 5. Additionally, we assume that the features available to the human and the ML model are denoted by $\mathbf{X}_{\text{H}} \in \mathbb{R}^{d_{\text{H}}}$ and $\mathbf{X}_{\text{M}} \in \mathbb{R}^{d_{\text{M}}}$ respectively, where d_{H} and d_{M} indicate the number of features available to the human and the machine respectively, with

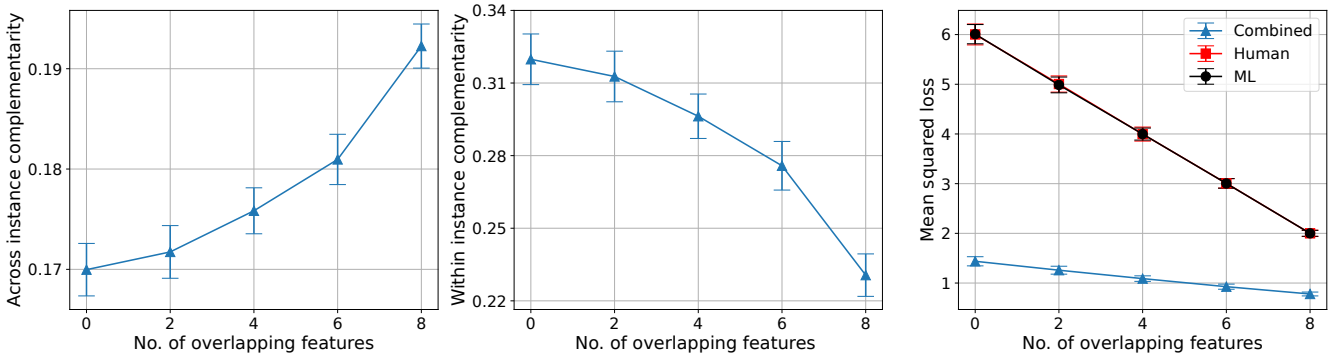

Figure 2: We plot the outcomes of Experiment I described in Section 5.1. The x-axis indicates the number of features that both the human and the ML model have access to. In each of the three figures, we plot an outcome metric for the optimal joint policy, namely across-instance complementarity (3), within-instance complementarity (4) and mean squared loss of the policy compared to the target outcome Y. The markers show the mean value and the error bars indicate the standard deviation, based on 200 iterations. On the x-axis, we skip $x = 10$, as it is a straightforward setting where both the agents have access to all the features, so there is no complementarity, $c_{\text{within}} = c_{\text{across}} = 0$. Note that all three plots have different ranges on the y-axis, with $c_{\text{across}} \in [0, 0.25]$, $c_{\text{within}} \in [0, 1]$. To read these plots, we focus on relative values within plots, and not on absolute values across plots. We observe that c_{across} increases while c_{within} decreases as the number of overlapping features increases. When the agents have no overlapping features ($x = 0$) the two agents have more likely to be equally benefitial for each decision leading to a higher within-instance complementarity. Meanwhile, when both have largely overlapping information ($x = 8$), the combination is more likely to show across-instance complementarity, the gains of going with the better decision-maker outweighing the possible gains from combination on each instance.

$d_{\text{H}}, d_{\text{M}} \leq d = 10$. Given the input information available to them, the human and the machine learn a policy using linear regression on the training data, given by $\pi_{\text{H}} : \mathbb{R}^{d_{\text{H}}} \to \mathbb{R}$ and $\pi_{\text{M}} : \mathbb{R}^{d_{\text{M}}} \to \mathbb{R}$ respectively. Using the optimization problem setup in (1) and (2), we conduct simulations to analyse the amount and type of complementarity achieved by the combination of human and ML agents with different information. Consequently, we conduct two sets of experiments.

Experiment I. We consider the setting where the human and ML have access to some common features and some non-common features as is typical of many real-world settings, as described in Section 3. Specifically, out of $d = 10$ features in our setting, the human and the ML both have access to z common features, and each has access to an additional $\frac{10-z}{2}$ features that only they can observe, where $z \in [d]$. We plot the outcomes of this experiment in Figure 2, where the x-axis of each plot indicates z (the degree of overlap between human and ML feature sets). Interestingly, we observe that while across-instance complementarity increases non-linearly with the number of overlapping features, within-instance complementarity decreases non-linearly. This suggests that when the two agents have access to many non-overlapping features, it would be important to use both the agents' decisions to come to a final decision on a given instance. On the other hand, in a setting with few overlapping features, the importance of collaboration on each instance reduces and it may be prudent to consider routing tasks to either the human or the machine for making the final decision. Furthermore, in the third plot, we observe that the combined decision has a strictly lower loss than either the human or the ML in isolation. Importantly, the gains achieved by the combined decision indicated by difference between the loss achieved by the individual agents and that by the combination is reducing as the number of overlapping features decreases. This suggests that depending upon the number of overlapping features and the resulting gain in accuracy, one may decide to forego joint human-ML decisions. We discuss this in more detail in Section 6.

Experiment II. Next, we consider a setting where the human has access to nine of the features $\mathbf{X}_{\text{H}} \in \mathbb{R}^9$ and the machine has access to the remaining tenth feature $\mathbf{X}_{\text{M}} \in \mathbb{R}$. Within this setting, we simulate the types of information asymmetry identified in Holstein et al. (2023). In this work on human-ML complementarity, the authors distinguish between non-overlapping features based on their "predictive power" which they define for any feature as the increase in training accuracy of a model as a result of including the feature. To simulate this, we vary the predictive power of the feature available to the ML model by introducing multiplicative random noise. Recall that $Y = \mathbf{X}^\top \beta + \epsilon$ where $\beta = (1 \cdots 1) \in \mathbb{R}^d$. Now, we define a variable α and let the data available to the ML model $\mathbf{X}_{\text{M}} \in \mathbb{R}$ be based on α as:

$$\mathbf{X}_{\text{M}} = \begin{cases} \mathbf{X}_{10} & \text{if Binomial}(\alpha) = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

In this manner, by varying α over the range $[0, 1]$, we vary the predictive power of $\mathbf{X}_{\text{M}}$. For $\alpha = 0$ we have $\mathbf{X}_{\text{M}} = 0$ constantly, implying zero predictive power, and for $\alpha = 1$ we have $\mathbf{X}_{\text{M}} = \mathbf{X}_{10}$, implying the highest predictive power under the setting assumed. We show the outcomes of different complementarity measures under this setting in Figure 3. Observe that in the first plot, the across-instance com-

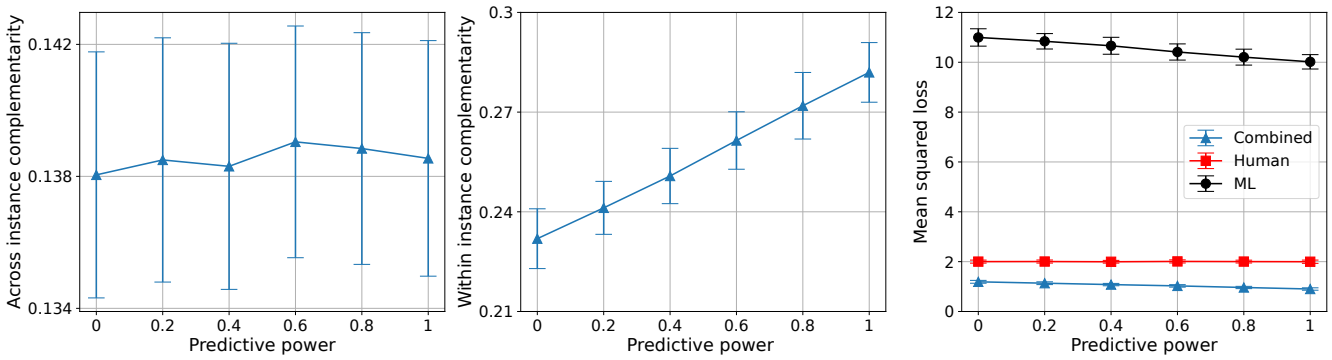

Figure 3: We plot the outcomes of Experiment II where the ML model has access to one feature and the human has access to the other nine features, described in Section 5.1. The x-axis indicates the predictive power of the feature $\mathbf{X}_{\mathrm{M}}$ that the machine has. In each of the three figures, we plot an outcome metric for the optimal joint policy, namely across-instance complementarity (3), within-instance complementarity (4) and mean squared loss of the policy compared to the target outcome Y. The markers show the mean value and the error bars indicate the standard deviation, based on 200 iterations. Note that all three plots have different overall ranges on the y-axis, with $c_{\text{across}} \in [0, 0.25]$, $c_{\text{within}} \in [0, 1]$. To read these plots, we focus on relative values within plots.

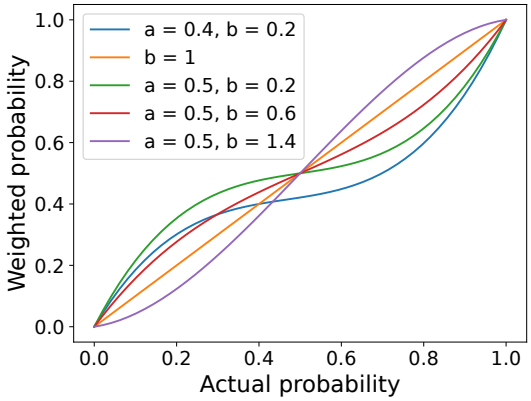

Figure 4: Examples of the probability weighting function used as the human's objective based on CPT. The x-axis specifies the actual probability and the y-axis indicates the perceived probability. Parameter a controls the fixed point and parameter b controls the curvature of the function. When $b < 1$, the probability weighting function has an inverted S-shape; when $b > 1$, the function has an S-shape.

plementarity does not change significantly with change in α. The reasoning behind this is the human has a large majority of the features, thus having a high contribution in the final decision for all settings of α. On the other hand, within-instance complementarity increases linearly with α, as increase in α implies that collaborating with the ML model on each instance will increase the predictive power of the overall policy. We also see that, as expected, the loss of the joint decision-maker improves as the predictive power increases.

5.2 Different Objective Functions

In this setting, the human and ML decision-makers have different objectives, which is a common source of complementarity in human and ML decision-making as noted in our taxonomy (Section 3). This may arise from the fact that ML models evaluate risks differently from humans. How agents evaluate the risks of an uncertain event is closely connected to how they perceive probabilities associated with this event. While ML models treat all probabilities according to their measured value, captured in their objective function as expected risk, humans tend to overweight small probabilities and underweight high ones, as suggested in Cumulative Prospect Theory (CPT) (Tversky and Kahneman 1992). To capture this in our simulation, we model the human's objective function incorporating CPT as described in Leqi, Prasad, and Ravikumar (2019b).

More specifically, while the ML model's objective is to minimize the expected value of the squared error, $F_{\mathrm{M}}(\pi_{\mathrm{M}}) = \frac{1}{n}\sum_{i=1}^{n}(\pi_{\mathrm{M}}(\mathbf{X}_i) - Y_i)^2$, the human's objective is to minimize $F_{\mathrm{H}}(\pi_{\mathrm{H}}) = \sum_{i=1}^{n}\frac{v_i}{n}(\pi_{\mathrm{H}}(\mathbf{X}_i) - Y_i)^2$ where v_i reflects how humans overweigh and downweigh certain probabilities. As illustrated in Figure 4, v_i is parameterized by two parameters $a \in [0, 1]$ and $b \in \mathbb{R}_+$ for specifying the fixed point and curvature of human's probability weighting function.[3] Notably, when $b = 1$, the probability weighting function becomes the identity function and v_i becomes 1 for all $i \in [n]$, suggesting that $F_{\mathrm{M}} = F_{\mathrm{H}}$. For a more detailed explanation on the relation among the parameters a, b, the probability weighting function, and the factor v_i in the objective function F_{H}, we refer the readers to Leqi, Prasad, and Ravikumar (2019b)[Section 3]. Lastly, we consider that the objective for the final decision balances between the human and the ML objective, defined as $F(\pi) = \theta F_{\mathrm{M}}(\pi) + (1-\theta)F_{\mathrm{H}}(\pi)$ where $\theta \in [0, 1]$ is a parameter controlling the overall objective function. By varying parameters θ, a and b, we inspect how the difference in objective functions of the two agents

[3] The exact form of v_i is defined using the derivative of the probability weighting function shown in Figure 4. More specifically, $v_i = \frac{3-3b}{a^2-a+1}(\frac{3i^2}{n^2} - \frac{2(a+1)i}{n} + a) + 1$.

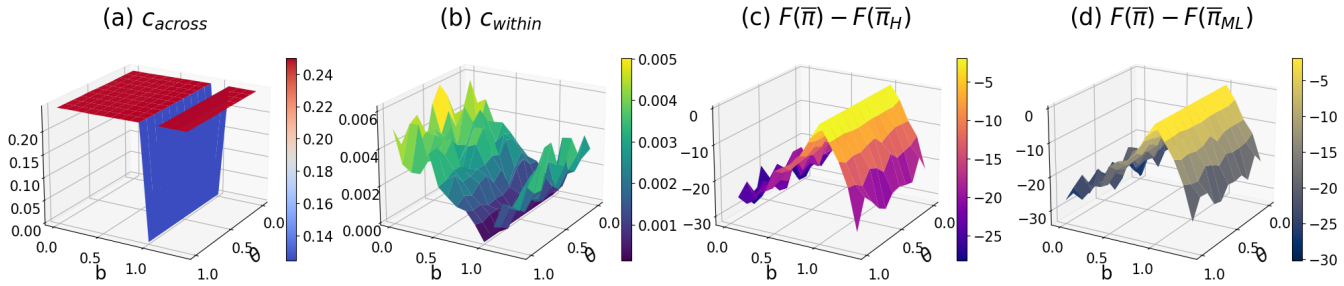

Figure 5: We plot the outcomes of the experiment where the ML and human have different objectives. For all plots, we set the probability weighting function parameter $a = 0.5$. The x-axis gives the b values, which specify the curvature of the probability weighting function; the y-axis gives the θ value, which specifies the overall objective function. In the first two plots, the z-axis shows the across-instance complementarity c_{across} and within-instance complementarity c_{within}, respectively. When $b = 1$ (i.e., $F_{\text{H}} = F_{\text{M}}$), both c_{across} and c_{within} reach their lowest values. We observe that c_{across} is high while c_{within} is low, indicating that the final decision of each task instance is more likely to rely on a single agent. In the last two plots, the z-axis shows $F(\overline{\pi}) - F(\overline{\pi}_{\text{H}})$ and $F(\overline{\pi}) - F(\overline{\pi}_{\text{M}})$, respectively. In both plots, the differences are below 0, suggesting that the joint policy performs better compared to π_{H} and π_{M} under the overall objective function F. All values are averaged across 5 seeds.

and the joint decision affects the amount and type of complementarity that can be achieved in this setting.

As observed in Figure 5 (c) and (d), the objective function differences $F(\overline{\pi}) - F(\overline{\pi}_{\text{H}})$ and $F(\overline{\pi}) - F(\pi_{\text{M}})$ remain below 0, suggesting that the learned joint policy outperforms both π_{H} and π_{M} under the overall objective function F. For both across-instance complementarity c_{across} and within-instance complementarity c_{within}, we find that when $b = 1$, i.e., when the human and machine objectives are the same, their values are the lowest and are around 0 (Figure 5 (a) and (b)). This is to be expected because when the overall objective is the same as that of the human and the machine, there is no complementarity. When $b \neq 1$, c_{across} is relatively high while c_{within} is rather low, suggesting that the optimal joint decision-maker does not need to rely on both agents for making a decision on most instances. Instead, a better form of collaboration between the human and the ML model is to defer each instance to one of the decision-makers. This is a somewhat unintuitive result since the overall objective function is a convex combination of the human's and the machine's, yet the final optimal decision is not. Importantly, this analysis shows evidence that we need to understand the mechanism of human-ML complementarity to inform how to design the best aggregation mechanism.

6 Discussion

Our work contributes a deeper understanding of possible mechanisms for complementary performance in human-ML decision-making. Synthesizing insights across multiple research areas, we present a taxonomy characterizing potential complementary strengths of human and ML-based decision-making. Our taxonomy provides a pathway for reflection among researchers and practitioners working on human-ML collaboration to understand the potential reasons for expecting complementary team performance in their corresponding application domains. Our hope is that the research community will use this taxonomy to clearly communicate their hypotheses about the settings where they expect human-ML complementarity in decision-making.

Drawing upon our taxonomy, we propose a problem setup for optimal convex combination of the human and ML decisions and associated metrics for complementarity. Our proposed framework unifies several previously proposed approaches to combining human-ML decisions. Critically, an analysis of our framework suggests that the optimal mechanism by which human and ML-based judgments should be combined depends upon the specific relative strengths each exhibits in the decision-making application domain at hand. Our optimization setup can be used to generate hypotheses about optimal ways of combining human and ML-based judgments in particular settings, as demonstrated by the simulations in Section 5. For this, one may use historical decision-making data or models of decision-making for the human and the machine agent. These simulations also help researchers and practitioners understand the trade-offs involved in implementing human-ML collaboration in a decision-making setting by comparing the potential gains in accuracy against the cost of implementation. It is worth noting here that while the joint decision-maker is a theoretical idealized version, in reality the accuracy of the joint decision-maker may be lower due to inefficiencies of real-world decision-making by a human. Thus, it would be useful to quantify the potential benefits of joint decision-making before implementation. Further, empirically testing the hypotheses and trade-offs presented by our simulations is of great theoretical and practical interest.

Finally, we invite extensions and modifications to our taxonomy, and hope that it serves as a stepping stone toward a theoretical understanding of the broader conditions under which we can and cannot expect human-ML complementarity. For example, we invite future research to explore extensions of our proposed optimization problem setup to contexts where predictions do not straightforwardly translate to decisions (Kleinberg et al. 2018), as well as to settings where the optimal combination of human and ML-based judgment cannot be captured through a convex aggregation function.

Acknowledgments

We thank the members of the FEAT ML reading group at Carnegie Mellon University and our anonymous reviewers for their insightful feedback that helped improve this work. During the course of this research, we were supported in part by the UL Research Institutes through the Center for Advancing Safety of Machine Intelligence (CASMI) at Northwestern University. CR was supported partly by the J.P. Morgan AI Research Fellowship, the IBM PhD Fellowship and NSF grant 1763734. LL was supported by the Open Philanthropy AI Fellowship. HH acknowledges support from NSF (IIS2040929 and IIS2229881) and PwC (through the Digital Transformation and Innovation Center at CMU). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not reflect the views of NSF or other funding agencies.

References

Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P.; Cao, X.; Khosravi, A.; Acharya, U. R.; et al. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*.

Adadi, A.; and Berrada, M. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6: 52138–52160.

Alkhatib, A. 2021. To live in their utopia: Why algorithmic systems create absurd outcomes. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–9.

Amitay, S.; Guiraud, J. A.; Sohoglu, E.; Zobay, O.; Edmonds, B. A.; xuan Zhang, Y.; and Moore, D. R. 2013. Human Decision Making Based on Variations in Internal Noise: An EEG Study. *PLoS ONE*, 8.

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. Accessed: 2023-06-01.

Bansal, G.; Nushi, B.; Kamar, E.; Horvitz, E.; and Weld, D. S. 2021. Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 11405–11414.

Bien, N.; Rajpurkar, P.; Ball, R.; Irvin, J.; Park, A.; Jones, E.; Bereket, M.; Patel, B.; Yeom, K.; Shpanskaya, K.; Halabi, S.; Zucker, E.; Fanton, G.; Amanatullah, D.; Beaulieu, C.; Riley, G.; Stewart, R.; Blankenberg, F.; Larson, D.; Jones, R.; Langlotz, C.; Ng, A.; and Lungren, M. 2018. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: Development and retrospective validation of MRNet. *PLoS Medicine*, 15(11). Publisher Copyright: © 2018 Bien et al. http://creativecommons.org/licenses/by/4.0/.

Bordt, S.; and von Luxburg, U. 2020. When Humans and Machines Make Joint Decisions: A Non-Symmetric Bandit Model. *arXiv preprint arXiv:2007.04800*.

Bottou, L. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, 177–186. Springer.

Brenner, L.; Griffin, D.; and Koehler, D. 2005. Modeling patterns of probability calibration with random support theory: Diagnosing case-based judgment. *Organizational Behavior and Human Decision Processes*, 97: 64–81.

Brown, A.; Chouldechova, A.; Putnam-Hornstein, E.; Tobin, A.; and Vaithianathan, R. 2019. Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-making in Child Welfare Services. In *CHI Conference on Human Factors in Computing Systems*, 41. ACM.

Bussmann, N.; Giudici, P.; Marinelli, D.; and Papenbrock, J. 2021. Explainable Machine Learning in Credit Risk Management. *Computational Economics*, 57.

Chouldechova, A.; and Roth, A. 2020. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5): 82–89.

De-Arteaga, M.; Fogliato, R.; and Chouldechova, A. 2020. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–12.

Donahue, K.; Chouldechova, A.; and Kenthapadi, K. 2022. Human-Algorithm Collaboration: Achieving Complementarity and Avoiding Unfairness. *arXiv preprint arXiv:2202.08821*.

Dressel, J.; and Farid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4(1): eaao5580.

Findling, C.; and Wyart, V. 2021. Computation noise in human learning and decision-making: origin, impact, function. *Current Opinion in Behavioral Sciences*, 38: 124–132. Computational cognitive neuroscience.

Finlayson, S. G.; Chung, H. W.; Kohane, I. S.; and Beam, A. L. 2018. Adversarial attacks against medical deep learning systems. *arXiv preprint arXiv:1804.05296*.

Fitzgerald, C.; and Hurst, S. 2017. Implicit bias in healthcare professionals: a systematic review. *BMC Medical Ethics*, 18.

Fogliato, R.; Chouldechova, A.; and Lipton, Z. 2021. The Impact of Algorithmic Risk Assessments on Human Predictions and Its Analysis via Crowdsourcing Studies. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).

Friedman, J. H. 2017. *The elements of statistical learning: Data mining, inference, and prediction*. springer open.

Gao, R.; Saar-Tsechansky, M.; De-Arteaga, M.; Han, L.; Lee, M. K.; and Lease, M. 2021. Human-AI Collaboration with Bandit Feedback. *arXiv preprint arXiv:2105.10614*.

Gentner, D.; and Stevens, A. L. 2014. *Mental models*. Psychology Press.

Gilpin, L. H.; Bau, D.; Yuan, B. Z.; Bajwa, A.; Specter, M.; and Kagal, L. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, 80–89. IEEE.

Gopnik, A.; and Wellman, H. M. 2012. Reconstructing constructivism: Causal models, Bayesian learning mechanisms, and the theory theory. *Psychological bulletin*, 138(6): 1085.

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

Green, B.; and Chen, Y. 2021. Algorithmic risk assessments can alter human decision-making processes in high-stakes government contexts. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2): 1–33.

Guerdan, L.; Coston, A.; Wu, Z. S.; and Holstein, K. 2023. Ground (less) Truth: A Causal Framework for Proxy Labels in Human-Algorithm Decision-Making. *arXiv preprint arXiv:2302.06503*.

Gulshan, V.; Peng, L.; Coram, M.; Stumpe, M. C.; Wu, D.; Narayanaswamy, A.; Venugopalan, S.; Widner, K.; Madams, T.; Cuadros, J.; Kim, R.; Raman, R.; Nelson, P. C.; Mega, J. L.; and Webster, D. R. 2016. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, 316(22): 2402–2410.

Hemmer, P.; Schemmer, M.; Kühl, N.; Vössing, M.; and Satzger, G. 2022. On the effect of information asymmetry in human-AI teams. *arXiv preprint arXiv:2205.01467*.

Hoffman, M.; Kahn, L. B.; and Li, D. 2017. Discretion in hiring. *The Quarterly Journal of Economics*, 133(2): 765–800.

Holstein, K.; and Aleven, V. 2021. Designing for human-AI complementarity in K-12 education. *arXiv preprint arXiv:2104.01266*.

Holstein, K.; Aleven, V.; and Rummel, N. 2020. A Conceptual Framework for Human–AI Hybrid Adaptivity in Education. In *Artificial Intelligence in Education*, 240–254. Cham: Springer International Publishing. ISBN 978-3-030-52237-7.

Holstein, K.; De-Arteaga, M.; Tumati, L.; and Cheng, Y. 2023. Toward supporting perceptual complementarity in human-AI collaboration via reflection on unobservables. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1): 1–20.

Hu, X.; Rudin, C.; and Seltzer, M. 2019. Optimal sparse decision trees. *Advances in Neural Information Processing Systems (NeurIPS)*.

Hüllermeier, E.; and Waegeman, W. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 1–50.

Jacobs, A. Z.; and Wallach, H. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 375–385.

Jarrahi, M. H. 2018. Artificial Intelligence and the Future of Work: Human-AI Symbiosis in Organizational Decision Making. *Business Horizons*, 61.

Kahneman, D.; Rosenfield, A. M.; Gandhi, L.; and Blaser, T. 2016. Noise: How to overcome the high, hidden cost of inconsistent decision making. *Harvard business review*, 94(10): 38–46.

Kawakami, A.; Sivaraman, V.; Cheng, H.-F.; Stapleton, L.; Cheng, Y.; Qing, D.; Perer, A.; Wu, Z. S.; Zhu, H.; and Holstein, K. 2022. Improving human-AI partnerships in child welfare: Understanding worker practices, challenges, and desires for algorithmic decision support. In *CHI Conference on Human Factors in Computing Systems*, 1–18.

Keswani, V.; Lease, M.; and Kenthapadi, K. 2021. Towards Unbiased and Accurate Deferral to Multiple Experts. *ACM Conference on Artificial Intelligence, Ethics, and Society*.

Khim, J.; Leqi, L.; Prasad, A.; and Ravikumar, P. 2020. Uniform convergence of rank-weighted learning. In *International Conference on Machine Learning*, 5254–5263. PMLR.

Kirwan, J. R.; de Saintonge, D. M. C.; Joyce, C. R. B.; and Currey, H. L. F. 1983. Clinical judgment in rheumatoid arthritis. I. Rheumatologists' opinions and the development of 'paper patients'. *Annals of the Rheumatic Diseases*, 42: 644 – 647.

Kleinberg, J.; Lakkaraju, H.; Leskovec, J.; Ludwig, J.; and Mullainathan, S. 2018. Human decisions and machine predictions. *The quarterly journal of economics*, 133(1): 237–293.

Koran, L. M. 1975. The Reliability of Clinical Methods, Data and Judgments. *New England Journal of Medicine*, 293(14): 695–701.

Kozyreva, A.; and Hertwig, R. 2021. The interpretation of uncertainty in ecological rationality. *Synthese*, 198(2): 1517–1547.

Kruppa, J.; Schwarz, A.; Arminger, G.; and Ziegler, A. 2013. Consumer credit risk: Individual probability estimates using machine learning. *Expert Systems with Applications*, 40(13): 5125–5131.

Lai, V.; Chen, C.; Liao, Q. V.; Smith-Renner, A.; and Tan, C. 2021. Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. *arXiv preprint arXiv:2112.11471*.

Lake, B. M.; Ullman, T. D.; Tenenbaum, J. B.; and Gershman, S. J. 2017. Building machines that learn and think like people. *Behavioral and brain sciences*, 40.

Leqi, L.; Prasad, A.; and Ravikumar, P. 2019a. On Human-Aligned Risk Minimization. In *Advances in Neural Information Processing Systems*.

Leqi, L.; Prasad, A.; and Ravikumar, P. K. 2019b. On Human-Aligned Risk Minimization. *Advances in Neural Information Processing Systems*, 32: 15055–15064.

Lipton, Z. C. 2018. The mythos of model interpretability. *Queue*, 16(3): 31–57.

Little, K. B. 1961. Confidence and Reliability. *Educational and Psychological Measurement*, 21(1): 95–101.

Lurie, E.; and Mulligan, D. K. 2020. Crowdworkers are not judges: Rethinking crowdsourced vignette studies as a risk assessment evaluation technique. *Proceedings of the Workshop on Fair and Responsible AI at CHI 2020*.

Madras, D.; Pitassi, T.; and Zemel, R. 2018. Predict responsibly: improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, 6147–6157.

Marr, D.; and Poggio, T. 1977. From understanding computation to understanding neural circuitry. *Neurosciences research program bulletin*, 15: 470–488.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267: 1–38.

Mitchell, S.; Potash, E.; Barocas, S.; D'Amour, A.; and Lum, K. 2018. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*.

Mozannar, H.; and Sontag, D. 2020. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, 7076–7087. PMLR.

Nisbett, R. E.; and Wilson, T. D. 1977. Telling more than we can know: verbal reports on mental processes. *Psychological review*, 84(3): 231.

Obermeyer, Z.; Powers, B.; Vogeli, C.; and Mullainathan, S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464): 447–453.

Okati, N.; De, A.; and Gomez-Rodriguez, M. 2021. Differentiable Learning Under Triage. *arXiv preprint arXiv:2103.08902*.

Patel, B.; Rosenberg, L.; Willcox, G.; Baltaxe, D.; Lyons, M.; Irvin, J.; Rajpurkar, P.; Amrhein, T.; Gupta, R.; Halabi, S.; Langlotz, C.; Lo, E.; Mammarappallil, J.; Mariano, A.; Riley, G.; Seekins, J.; Shen, L.; Zucker, E.; and Lungren, M. 2019. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ Digital Medicine*, 2(1). Publisher Copyright: © 2019, The Author(s).

Quiñonero-Candela, J.; Sugiyama, M.; Lawrence, N. D.; and Schwaighofer, A. 2009. *Dataset shift in machine learning*. Mit Press.

Raghavan, M.; Barocas, S.; Kleinberg, J.; and Levy, K. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 469–481.

Raghu, M.; Blumer, K.; Corrado, G.; Kleinberg, J.; Obermeyer, Z.; and Mullainathan, S. 2019. The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220*.

Rajpurkar, P.; O'Connell, C.; Schechter, A.; Asnani, N.; Li, J.; Kiani, A.; Ball, R.; Mendelson, M.; Maartens, G.; Van Hoving, D.; Griesel, R.; Ng, A.; Boyles, T.; and Lungren, M. 2020. CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. *NPJ Digital Medicine*, 3: 115.

Rastogi, C.; Zhang, Y.; Wei, D.; Varshney, K. R.; Dhurandhar, A.; and Tomsett, R. 2022. Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making. *ACM CSCW*.

Roth, E. M.; Sushereba, C.; Militello, L. G.; Diiulio, J.; and Ernst, K. 2019. Function Allocation Considerations in the Era of Human Autonomy Teaming. *Journal of Cognitive Engineering and Decision Making*, 13(4): 199–220.

Russakovsky, O.; Li, L.-J.; and Li, F.-F. 2015. Best of both worlds: Human-machine collaboration for object annotation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition*, 2121–2131.

Saxena, D.; Badillo-Urquiola, K.; Wisniewski, P. J.; and Guha, S. 2020. A Human-Centered Review of Algorithms used within the US Child Welfare System. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15.

Shrestha, Y. R.; Ben-Menahem, S. M.; and Von Krogh, G. 2019. Organizational decision-making structures in the age of artificial intelligence. *California Management Review*, 61(4): 66–83.

Simon, H. A. 1979. Rational decision making in business organizations. *The American economic review*, 69(4): 493–513.

Steyvers, M.; Tejeda, H.; Kerrigan, G.; and Smyth, P. 2022. Bayesian modeling of human–AI complementarity. *Proceedings of the National Academy of Sciences*, 119(11): e2111547119.

Tan, S.; Adebayo, J.; Inkpen, K.; and Kamar, E. 2018. Investigating Human+ Machine Complementarity for Recidivism Predictions. *arXiv preprint arXiv:1808.09123*.

Tenenbaum, J. B.; Kemp, C.; Griffiths, T. L.; and Goodman, N. D. 2011. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022): 1279–1285.

Trönnberg, C.-C.; and Hemlin, S. 2014. Lending decision making in banks: A critical incident study of loan officers. *European Management Journal*, 32(2): 362–372.

Tschandl, P.; Codella, N.; Halpern, A.; Puig, S.; Apalla, Z.; Rinner, C.; Soyer, P.; Rosendahl, C.; Malvehy, J.; Zalaudek, I.; Argenziano, G.; Longo, C.; and Kittler, H. 2020. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26.

Tversky, A.; and Kahneman, D. 1992. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5: 297–323.

Wilder, B.; Horvitz, E.; and Kamar, E. 2020. Learning to complement humans. *arXiv preprint arXiv:2005.00582*.

Wistrich, A. J.; and Rachlinski, J. J. 2017. Implicit Bias in Judicial Decision Making How It Affects Judgment and What Judges Can Do About It. *Chapter 5: American Bar Association, Enhancing Justice*.

Zhang, H.; and Maloney, L. 2012. Ubiquitous Log Odds: A Common Representation of Probability and Frequency Distortion in Perception, Action, and Cognition. *Frontiers in Neuroscience*, 6: 1.