CAREER: Understanding and Accounting for Human Behavior in Machine-in-the-Loop Decision Making

1 Introduction

Machine learning (ML) has gained significant progress in the past decade. With the improving predictive power, ML has been increasingly involved in various decision making in our daily life, from determining which movies to recommend, to who to approve a loan application, to whether to give bail to a defendant. While fully automated decision making is the goal for tasks in some domains, such as autonomous driving, in many other tasks, we should not, or do not want to, delegate decision making entirely to ML. For example, when the stake of the task is high and the objectives of the task are hard to be accurately specified, such as in government policy making or military applications, fully automatic decision making using ML may lead to suboptimal outcomes and has not achieved the level of being fully trusted. In these cases, humans are often brought into the loop to make the final call on what decisions to take. In addition, for tasks that involve human preferences or enjoyments, such as in deciding which restaurants to go to or which destinations to travel to, while ML might be able to provide useful information/recommendations, humans are still naturally the final decision makers in these tasks.

These considerations lead to a new paradigm of *decision making with machines in the loop*, where ML provides information to humans, who can then incorporate the information to make the final decisions. Here, the goal of machine learning is to *augment*, instead of *replacing*, humans in decision making. This type of machine-in-the-loop decision-making paradigm has been emerging in a variety of domains. For example, online users receive product recommendations from online platforms to decide which product to purchase. Doctors utilize predictions from ML to decide the treatments for patients. Judges use risk assessment tools to evaluate the recidivism risk of a defendant to make the bail decisions. Meanwhile, while machine-in-the-loop decision making presents significant promises, there are challenges as well. In particular, humans are known to exhibit behavioral biases in making decisions. How do we take into account of human behavior in designing ML algorithms to assist humans? Moreover, there might be multiple objectives to balance and trade-off during decision making, how should we design the assistive ML algorithms that balance objectives such that it aligns with human values?

In this research proposal, I plan to investigate machine-in-the-loop decision making, with a focus on studying the *information exchange* between ML and humans. The goal is to **utilize machine learning to assist humans in making better decisions while taking into account human behavior and preferences**. This research is interdisciplinary in nature, requiring techniques and insights from machine learning, optimization, algorithmic economics, and online behavioral social sciences. The proposed research combines both theoretical and empirical approaches: new theories will be developed to incorporate human models into decision making frameworks, and empirical experiments will be conducted to better understand and model human behavior. More specifically, I will investigate the following three research thrusts:

- **Developing algorithms for information design in machine-in-the-loop decision making**: We will develop algorithmic frameworks for ML to assist humans in making better decisions. In particular, we will explore the usage of information design, i.e., how ML can present information to humans such that humans can make better decisions. The information can be structured either as a recommendation of action, or a conditional distribution of outcomes when a certain action is taken. In addition to study optimal information design when human behavior models are known, we will also investigate settings in which human behavior is unknown and we need to either learn from past interactions or have an information design that is robust to a range of possible behavior.
- **Understanding and modeling humans in machine-aided decision making**: In computational frameworks with humans in the loop, humans are often assumed to be *Bayesian rational* when making deci-

sions, i.e., they process information in a Bayesian manner and take actions that maximize their expected payoff. However, as empirically observed in psychology and behavioral economics, humans often exhibit systematic biases in processing information and making decisions. We aim to examine human decision making in a variety of settings by conducting a series of large-scale behavioral experiments and develop more realistic human models that aligns with empirical observations. Moreover, since the goal of human models is to accurately predict what humans will do, we will develop a framework based on machine learning theory to discuss the expressiveness of different behavior models.

- **Leveraging human decision-makers in information design**: In the above discussion of formulating machine-in-the-loop decision making as a constraint optimization problem, we have abstracted away of what the objective of the optimization should be. This poses a potential concern that, while the goal of machine-in-the-loop decision making is to assist humans in making decisions, an ill-defined problem formulation could lead to undesired outcome. In this thrust, we plan to borrow ideas from participatory design for algorithmic governance [63, 4, 76, 19] and include human decision-makers in the loop in shaping the formulation of the information design problem.

Long-term Goal. My career goal is to develop the foundations for humans and ML to collaborate together and solve problems neither can solve alone. This requires the advancements of machine learning, the understanding of humans, and the utilization of their interactions. This research proposal serves as the stepping stone to achieve that goal by investigating how to design machine learning algorithms to assist humans in making better decisions while taking into account human behavior.

Intellectual Merit. This proposed research will contribute to the empirical understanding of human behavior in ML-assisted decision making. It will also provide theoretical foundations for studying the interactions of humans and learning algorithms, through incorporating human models in computational frameworks. The results of the proposal will provide insights on developing human-centered machine learning algorithms and in combining humans and machines to solve problems neither can solve alone. This research is interdisciplinary in nature, combining ideas and techniques from machine learning, algorithmic economics, and online behavioral social science.

PI Qualifications. The PI has extensive research experience in the design and analysis of systems with humans in the loop. From the machine learning perspective, the PI has explored the problem of learning from noisy data contributed by humans and optimally matching humans with suitable tasks [43, 46]. The PI also designed data elicitation mechanisms to enable more efficient learning [1, 49, 44]. From the economics perspective, the PI has explored the design of different types of incentives, such as reputation [45], monetary payments [47, 50], and attention [65]. Moreover, the PI has explored behavioral aspects of humans in computational environments [91, 97] and address ethical considerations [95, 94]. In addition to the theoretical and algorithmic studies, the PI has experiences in conducting large-scale online behavioral experiments to understand human behavior, such as how users react to financial incentives in crowdsourcing markets [48] and how users are influenced when interacting with other users [93, 30], and how users make decisions under uncertainty [92]. The PI is active in the research communities of machine learning, algorithmic economics, and human computation, including organizing workshops at NIPS and HCOMP to explore the connection of crowdsourcing and machine learning and to foster the study of theoretical foundations of human computation. The PI also served as the Works-in-Progress and Demonstration Co-Chair of HCOMP 2019, the premier conference in the study of human computation.The PI has served on the senior program committee or area chair of AAAI, IJCAI, and NeurIPS and the program committees (or the equivalent) of major conferences.

2 Background

This proposal aims to understand and address human behavior in decision making with machines in the loop. Below we provide a short background on machine-in-the-loop decision-making frameworks and also discuss human behavioral models.

2.1 Machine-in-the-Loop Decision Making

As machine learning gets more and more involved in decision-making our everyday life, it is of paramount importance to study how machine predictions impact human decision-making across a broad range of contexts. A related body of work in studying this human-machine interaction is human-in-the-loop machine learning [109, 106], in which computational systems rely on human involvements (such as labeling photos and correcting errors) to overcome limitations and improve their performance. Different from human-in-the-loop machine learning, where the goal is to incorporate humans to improve machine learning performance, in this proposal, we aim to study machine-in-the-loop decision making, in which the goal is to utilize machine learning to augment and assist humans in making decisions.

This human-centered focus in human-machine interactions has spurred various research themes in the recent years, including improving the overall performance of human-AI partnerships [37, 38, 61, 7, 62] and also in investigating the interpretability [82, 69, 40, 80, 58, 78] and trustworthiness [26, 32, 27, 68, 107, 108] of machine learning, which highlights the questions on how humans take the machine learning predictions to make decisions. In this research proposal, we aim to formalize the information exchange between humans and machines. Our goal is to study how humans process the machine-provided information make decisions accordingly and how machines can design information to assist or influence humans in making decisions.

2.2 Human Behavioral Models in Decision Making

Existing approaches in modeling humans in computational frameworks mostly fall into two categories. The first one involves the perspective of considering humans as data contributors. In this category, humans are often modeled as data sources that perform as independent, random variables that output data according to a generative model. This assumption enables the whole line of research works in the literature on label aggregation and truth discovery to incorporate human data in machine learning [24, 81, 18, 52, 103, 23, 110, 17]. In the second category, when we consider humans take actions to respond to the environment, humans are often assumed to be *Bayesian rational* decision makers aiming to take actions that maximize their expected utility [102, 10, 9, 42, 59, 2]. While these models provide elegant and simple formulations, they do not always capture true human behavior in the field.

This research proposal aims to understand human behavior on the decision-making perspective. Therefore, we plan to examine alternative models in the two stages of machine-in-the-loop decision making: 1) belief updating: how humans process the ML-provided information and update their beliefs on the state of the world and 2) decision making under uncertainty: how humans make decisions with their beliefs. From the belief updating perspective, while Bayesian models have been the prominent model in algorithmic works [98, 39, 36], it has also been consistently and widely observed in empirical studies that humans often deviate from being Bayesian [53, 100, 5, 89, 67, 72]. While there have been some alternative models in how humans process information to form their beliefs [74, 86, 70, 105, 83, 79], they are not widely adopted in algorithmic frameworks. From the decision-making perspective, the common assumption is expected utility theory [102] which assumes humans take actions to maximize their expected utility. There is again a substantial body of work in behavioral economics in studying the systematic deviations of human behavior from expected utility theory. For example, it is consistently observed that humans often over-estimate small probabilities (e.g., partly explaining why people buy lotteries despite its negative expected reward) and react more strongly to losses than gains. The most important theory that summarizes these systematic biases is perhaps the Nobel-winning *prospect theory* by Kahneman and Tversky [55]. Another commonly used theory (also Nobel-winning) is the discrete choice model [71, 87, 99], which accounts for the inherent randomness of human decision making by incorporating noises in the utility. In this proposal, we will conduct behavioral experiments to examine these models (and potentially additional models) in the context of machine-in-the-loop decision making.

3 Proposed Research

The proposed research is concerned with investigating human-ML partnership through focusing on human behavior in machine-in-the-loop decision making. We plan to empirically examine and model human behavior in the context of ML-assisted decision making and also develop algorithmic frameworks in designing information from ML to assist humans. In addition, we plan to investigate the potential negative impacts brought up by including machines in the loop of decision making and study approaches to mitigate the negative impacts.

3.1 Thrust 1: Designing Information in Machine-in-the-Loop Decision Making

In this thrust, we plan to develop algorithmic frameworks for machine learning to assist humans in making better decisions. As the central theme of this research proposal, we will explore the usage of *information design*, i.e., what types of information disclosing policy should ML choose to better assist humans. Consider, for example, ML-assisted navigation, the goal of ML is to parse and provide traffic information to human drivers so human drivers can decide on the best route to take. Our focus will be on designing the information structure to present to humans. The information can be structured either as a recommendation of action (e.g., which route to take) or a distribution of signals conditional on the realization of the world state (e.g., the estimated time for the particular route given the traffic). We also note that while the presentations of information, such as layout, color, or font size, language usages, are also important aspects of information design, they are not the focus of this proposed research.

To formulate this information design framework, we will start with the full information setting in which we assume human behavior models are known. This enables us to develop a game theoretical framework and formulate the information design as a bi-level optimization problem, where ML is choosing the optimal information design while considering humans optimizing their actions conditional on the provided information. We will then relax the full-knowledge assumption and investigate the associate learning and robust design problems.

3.1.1 Preliminary work: A Stackelberg game formulation

The information design framework can be formulated as a *Stackelberg game*, in which ML first decides on the information disclosing policy, and humans decide on what decision to take based on the provided information. My prior works have utilized the formulation of Stackelberg games in a range of different application domains. In particular, I have studied problem of contract design [47, 50], in which the firm first posts a contract and the employee/worker decides on the amount of effort/work in response to the contract, the problem of learning with strategic responses [97], in which the learner first posts a decision rule and the agent responds with the goal of receiving a favorably treatment, and Bayesian persuasion [28, 92], in which the sender decides an information disclosure strategy to persuade the receiver to take certain actions. These prior works will serve as the foundation to address the research questions in this thrust.

In particular, consider my most relevant work in the persuasion setting [28, 92], which aligns well with our objective if we consider the sender as ML and the receiver as human decision makers. We show that the information design problem in persuasion can be treated as a constrained optimization problem. Slightly more formally, let the state of the world be θ which is drawn from a finite set Θ according to a prior distribution $\mu_0 \in \Delta(\Theta)$. Let τ be the information scheme the sender chooses. Upon receiving the realization of the information scheme, the receiver can choose an action a from an action set A. The receiver's model is specified by $\mu(\mu_0, \tau)$, the posterior distribution induced by the signal scheme, and $P(a|\mu(\mu_0), \tau))$, the receiver's decision function, characterized by a distribution of decisions given the posterior. Let $V(a, \theta)$ be the sender's utility when the receiver takes action a and the state of the world be θ. Then the sender's information design problem can formulated as follows:

$$\max_{\tau} \ \mathbb{E}_{\theta}[\sum_{a} P(a|\mu(\mu_0, \tau))V(a, \theta)] \tag{1}$$

$$\text{s.t. } \ \mathbb{E}_{\theta,\tau}[\mu(\mu_0, \tau)] = \mu_0 \tag{2}$$

The objective corresponds to maximizing the sender's expected utility, while the constraint specifies that the receiver's belief updating needs to be *plausible*, i.e., the expectation of the posterior needs to be consistent with the prior. For different human behavioral models, we need to specify $\mu(\mu_0, \tau)$ to reflect how humans update their beliefs when being provided information and $P(a|\mu(\mu_0), \tau))$ to reflect how humans make decisions under uncertainty. My prior works have addressed the above optimization problem in settings with Bayesian rational receivers [28] and settings with Non-Bayesian-rational receivers [92], in which we assume the receiver updates her belief with probability weighting [71, 87, 99] on the prior and make decisions following the discrete choice model [105, 83, 79].

3.1.2 Task 1.1: Develop a framework for optimizing information with different human models

The preliminary results have set up the persuasion setting as a Stackelberg game and formulated information design as a constrained optimization problem. Again, note that the persuasion problem has a direct connection to our target setting of machine-in-the-loop decision making, if we map the ML as the sender who decides on what information to present and humans as receivers that decide on what decisions to make based on the provided information. Our preliminary results have focused on two specific human behavioral models, and each of them has led to very different set of results. For example, consider the case that humans are Bayesian rational, the decision function $P(a|\mu(\mu_0), \tau))$ is a delta function that puts all the probability mass on the action that maximize the receiver's payoff. When putting this decision function back to the optimization problem, the objective is non-continuous and the optimization is in general NP-hard to solve. On the other hand, when we consider the discrete choice model (let us abuse the notation and set $\mu = \mu(\mu_0, \tau)$), the decision function is in the form of: $P(a|\mu) = \frac{\exp\left(\beta \widehat{u}^R(a|\mu)\right)}{\sum_{a'} \exp\left(\beta \widehat{u}^R(a'|\mu)\right)}$, which is essentially a continuous softmax function. With this human behavioral model, the information design problem can be formulated as a convex optimization problem, and there exist efficient algorithms to find the optimal information design when the space of the information design is small.

The above discussion highlights the need to understand how different human behavior models impact the problem of optimal information design. In this task, we will assume the knowledge of human behavioral models and address the corresponding optimization problem. Depending on the human models, there will be two types of optimization problems to be addressed:

- Optimization with non-continuous objective: When the human decision model follow the expected utility theory or prospect theory (and possibly other variants), since human decision making will be in the form of choosing an action that maximizes the (possibly distorted) payoff function, the objective of the optimization problem will be non-continuous, and we cannot directly apply the standard first-order methods to solve the optimization problem. In this type of problem, we plan to utilize the techniques from recent research efforts in algorithmic persuasion [31, 33, 6] to characterize the equilibrium solution and the computational complexity. On a high-level, this line of approach often involves utilizing the duality theory to characterize the properties of the optimal solution. The characterizations can help reduce the search space for optimal solutions and make the optimization more efficient.

- Optimization with continuous objective: When the human decision model follow the discrete choice model or other models that lead to stochastic decision making, the optimization objective can usually be written down as a continuous differentiable function. This enables the first-order optimization techniques, such as gradient descent, to be applied in this optimization problem. In this type of problems, we plan to characterize the computational complexity and convergence to the optimal solution with different human models of belief updating and decision making. My prior work on the study of complexity and convergence rate of secure convex optimization [96] will serve as the technical foundation for addressing this problem.

3.1.3 Task 1.2: Design Information with uncertain human behavior

In the previous task, we start our investigation by assuming full knowledge of human behavior. While this assumption might be able to be approximately satisfied when we have an abundant of data of human behavior in the environment, it is generally a strong assumption and might not always be satisfied. In this task, we propose to address unknown human behavior is through learning from past interactions. Assume ML can sequentially interact with humans in decision making. We can infer humans' behavior models by utilizing the interaction of previous rounds (e.g., by observing human decisions on the past presented-information) and then update the information disclosure strategy in the future rounds. More formally, we can formulate this as a multi-armed bandit problem [60, 3, 11], with each possible information disclosure strategy as an arm. While the bandit formulation provides a nice foundation to perform exploration-exploitation trade-off (i.e., the trade-off of learning the effectiveness of an information strategy vs. deploying the strategy to obtain payoffs), the main challenge of our setting is that there are infinitely many information disclosure strategies (infinitely many arms), and standard bandit algorithms would not converge to the optimal policy since there are too many arms to explore.

We plan to address this challenge using the technique in my prior work on adaptive contract design [50], in which we aim to find the optimal contract to crowd workers with unknown cost/effort levels by adaptively updating the contracts and observe their performance. The key intuition of our proposed technique is that, when we select a contract, the response we obtained from humans not only provide us information about the posted contract but also the similar contracts (i.e., worker performance should be similar with similar payments). Therefore, we can propagate this information to nearby arms and achieve near-optimal learning. We plan to apply similar idea in information design through mapping the information disclosure policy to the contract. We would also discuss how different human behavior models, while unknown to us, would impact the learning efficiency.

3.1.4 Task 1.3: Robust information design

We plan to also study settings in which we cannot learn from past interactions, and the goal is to design an information policy that is *robust* to a set of candidate human models. We plan to address this problem by borrowing ideas from robust contract design [13, 22, 73, 15, 14, 25, 41, 16], in which robustness is defined as the worst-case optimal mechanisms, considering all the possible (unknown) actions players can take. We will start by building a connection between a *contract* in contract design and a *information strategy* in our setting. More specifically, since the main challenge here is due to this non-quantifiable uncertainty about humans, we can similarly define robustness based on the worst-case guarantee of human decisions (induced by unknown human behavior models), over all possible decisions humans might take. Our goal is to design *robust optimal* information strategy: the strategy is *robust optimal* if the worst case performance is (weakly) better than that of all other possible strategies. My prior work [97] on robust learning, which also utilizes the techniques from robust contract design to design robust decision rules for strategic users, will serve as the technical foundation for this proposed research.

3.2 Thrust 2: Understanding and Modeling Humans in Machine-Assisted Decision Making

In this research thrust, we plan to examine human behavior in the context of ML-assisted decision making, where humans receive and incorporate information from ML to make decisions. We plan to conduct large scale behavioral experiments in a wide range of settings. While there have been empirical works in psychology and cognitive science in examining human behavior in general decision making, as briefly surveyed in Section 2.2, these works are mostly smaller-scale lab studies and are not in the context of machine-in-the-loop decision making. Human decisions might deviate from these results when they are aware that the information is provided by ML. Therefore, we plan to conduct larger-scale experiments in the setting with machine-in-the-loop decision making to understand and model human decision-making with ML assistance across a wide range of context. To achieves this goal, we will leverage online experimental platforms, such as Amazon Mechanical Turk, for our experiments since these online platforms provide a natural solution to examine scenarios with machines in the loop and to engage a larger population for behavioral studies.

Our focus of the study will be on examining whether humans are Bayesian rational, the common assumption in the literature, when machine is in the loop, and if not, how humans deviate and how we can develop alternative models that better explain human behavior. We will also explore whether we can take interventions during human decision making to make them more towards being Bayesian rational (the *optimal* way of decision making from the objective perspectives). In addition, since the goal of using behavioral models to explain human behavior is to be able to predict human behavior for future scenarios. This aligns with the goal of machine learning, i.e., have low generalization error for unseen data points. Therefore, we will utilize this connection and examine the expressiveness of alternative behavioral models using the lens of machine learning theory.

3.2.1 Preliminary work

My prior works have addressed the questions of modeling real-world human behavior through conducting a series of online behavioral experiments, including examining whether crowd workers are rationally responding to monetary payments [48], whether and how human opinions are influenced when communicating with others [93, 30], and whether humans are Bayesian rational in the persuasion setting [92]. Here I describe my most relevant work [92] that serves as the starting work of the following proposed research. In particular, we focus on the persuasion setting [57, 56, 8] in which a sender aims to design information to persuade a receiver to take certain actions. We empirically examine that given prior beliefs, how the receivers (i.e., humans) process information to form their posterior and take actions.

Experiment design: We recruited 400 workers from Amazon Mechanical Turk. Each worker is asked to complete 20 questions, in which each question asks humans to perform a probabilistic-inference and decision-making task. In each question, workers are informed that there are two urns, Urn X and Urn Y, corresponding to two world states, where each of them contains certain fraction of red balls and blue balls (announced and known to workers). At the beginning of each question, an urn will be firstly secretly drawn according to the prior distribution announced to workers. After an urn is drawn, a ball will be drawn uniformly at random from this urn. The color of the drawn ball will be disclosed to the worker. Upon seeing the ball color, the worker is asked to guess on which urn is drawn, and she will get bonus for making the correct guess. This experiment setup aims to capture human decision-making process, including how humans update their prior beliefs (the prior of urn drawing) with additional information (realized ball drawn according to the commonly known ball compositions in urns) and make decisions (guessing which is the real urn).

To examine whether users are Bayesian rational, we conducted randomized experiments with two treatments, which differ in the prior distribution of the state. In *high prior* treatment, we fixed the prior to be $(0.4, 0.6)$ for urn X and urn Y, while in *low prior* treatment, the prior is fixed as $(0.2, 0.8)$. We then design eight ball compositions in urns (corresponding to signal structures) such that, conditional on the realization of a red ball draw, the Bayesian posterior would be $(0.2, 0.3, \dots, 0.9)$ for both treatments.

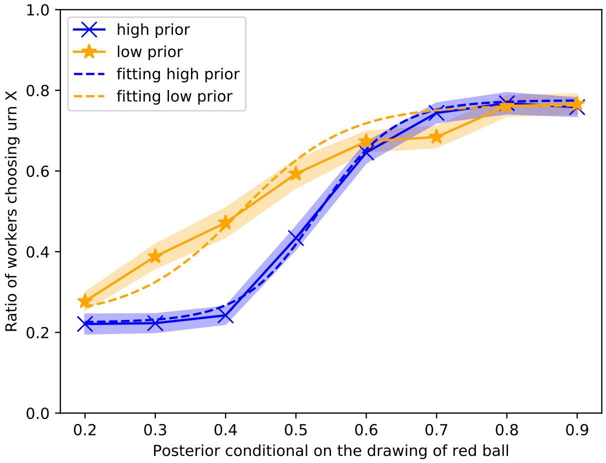


Figure 1: The solid lines represent the percentage of workers that choose Urn X conditional on a red ball realization. Shaded regions correspond to standard errors. Dashed lines correspond to fitted models using discrete choice model and probability weighting. The differences between the two treatments indicate that workers are not Bayesian and the sigmoid curve within the same treatment indicate that workers are not expected utility maximizers.

Experiment results: According to our experiment design, if workers update beliefs in a Bayesian manner, we should see no difference across two treatments. If workers are rational, we should see workers choosing urn X when their posterior is greater than 0.5 for both treatments. However, the results, as shown in Figure 1, demonstrate that worker behavior has significantly deviated from the model of Bayesian rationality. In particular, the differences between the two treatments demonstrate that workers are not updating their beliefs in a Bayesian manner, since their actions depend not only on the Bayes posterior but also on the prior. The sigmoid-shape curve for both treatment demonstrate that workers are not expected utility maximizer. We also show that, if we use an alternative human model (discrete choice model [71, 87, 99] coupled with probability weighting [105, 83, 79]) to fit the collected data, we can see that this alternative model provides a better explanation of workers' behavior

3.2.2 Task 2.1: Understand and model human behavior via behavioral experiments

Our preliminary results demonstrate that humans may not be Bayesian rational in a particular setting. In this task, we will conduct a series of large-scale behavioral experiments to examine human behavior and investigate when existing behavioral models explain well human behavior. We later will also consider cases when classical models fail and new models might need to be developed (see Section 3.2.4).

Candidate behavioral models: We will start our investigation by examining the classical theories in different scenarios and different information structure. These theories can be categorized into two groups: belief updating with new information and decision making under uncertainty.

In belief updating, let $P(e)$ denote the prior of event e and I denote the signal/information. The common Bayesian assumption states that human form the posterior given information I following the Bayes rule $P(e|I) = \frac{P(I|e)P(e)}{P(I)}$, which is a strong assumption requiring perfect reasoning. One alternative general framework to incorporate human biases in belief updating is to assume human posterior is in the form

$$P(e|I) \propto P(I|e)^{\alpha} P(e)^{\beta} P(I)^{\gamma} \tag{3}$$

This formulation captures various human biases (e.g., confirmation bias, anchoring effect) that weight too much on the prior information or on the additional signals through varying on the parameters of α, β, and γ.

In decision making under uncertainty, let $(p_1, x_1, ..., p_K, x_K)$ be the *prospect* of an action, where p_k represents the probability of the outcome x_k happens after taking the action. Let $v(x_k)$ represent the utility of the outcome x_k. There are a couple of main theories in explaining human decisions:

- Expected utility theory [102]: it predicts that humans will take the action that maximizes $\sum_{k=1}^{K} p_k v(x_k)$.
- Prospect theory [55]: it predicts that humans will take the action that maximizes $\sum_{k=1}^{K} \pi(p_k) u(v(x_k))$, where $\pi(\cdot)$ and $u(\cdot)$ models the humans' distorted interpretations on the probability and utility measure.
- Discrete choice model [71, 87, 99]: it incorporates the intrinsic randomness of human decision making and adds independently drawn noise ϵ to the expected utility, which leads to a stochastic decision process.

Experiments. We will conduct a series of large-scale behavioral experiments to examine which of the above theories aligns better with human decision making. The experiments will be conducted in a wide range of scenarios/tasks and under different information design (e.g., prior info representation and signal compositions) to examine the factors that influence the performance of each model. For the experiment designs, the **independent variables** we plan to control across different treatments are different levels of *prior* and *information* (following the composition of Equation 3) to measure the impacts of belief updating and on different amount of payments to understand whether humans are more likely to be Bayesian rational when the stake is higher. In addition, we will examine different setup of the case scenarios and information structure. For the **performance measure**, since each of the decision theory gives a prediction on what humans will do, we can measure the *empirical prediction accuracy* for each of the models to examine which model aligns with the real human decisions. Note that since our goal of having a behavioral model is to predict human behavior in future scenarios, this naturally corresponds to the generalization error in machine learning. We will discuss whether we can quantify the expressiveness of each behavioral model and new models are needed by drawing the connection to machine learning theory (see Section 3.2.4).

Expected outcome. This task will provides us an empirical understanding of how accurately each human decision model is in explaining real human decision-making and identify the key parameters (of the job at hand or the information structure) that impact the model performance under a wide spectrum of scenarios.

3.2.3 Task 2.2: Design interventions to induce rational decision making

The above task aims to understand and model real-world human behavior when responding to ML-provided information. Note that since being Bayesian rational is the *optimal* way (in an objective sense) of making decisions, another potentially equally important question is whether we can nudge humans to be more Bayesian rational. In this task, we plan to study interventions to nudge humans to be more Bayesian rational when making decisions. We plan to base our design on the well-celebrated dual process theory (DPT) [34, 54, 20]. DPT specifies two processes through which thoughts may arise—Type 1 processing and Type 2 processing. Type 1 processing is fast, automatic, instinctive, and unconscious, and Type 2 processing is slower, deliberate, rule-based, and conscious. While human usually utilize some combination of both intuitive (i.e., Type 1) and analytical (i.e., Type 2) processing during their decision making, it is believed that the default processing mode human brains would select is Type 1 processing. However, Type 1 processing is largely associated with the use of heuristics, thus excessive reliance of Type 1 processing would override Type 2 processing, trigger bias from humans, and lead to insufficient deliberation and unexamined decisions. Moreover, the risk of overusing Type 1 processing is particularly high when humans suffer from fatigue, sleep deprivation, and cognitive overload [20].

Design space of interventions. Based on DPT, a premise for nudging humans to be more Bayesian rational is to enable them to decouple from their own automatic responses in decision-making that are resulted from Type 1 processing, i.e., mitigate their biases in decision making. In other words, the key is to have humans actively engage in Type 2 processing and override Type 1 processing as needed. Addressing this key requirement, concrete procedures have been developed for cognitive sciences [88, 104], which includes a series of steps from raising humans' awareness of bias and motivating humans to correct bias, to enabling humans to use situational cues to recognize the need of debiasing, to instructing humans to inhibit heuristic responses and analyze alternative solutions. We can categorize the interventions based on when they appear: (1) *Pre-decision*: Interventions used at the pre-decision stage could serve two main goals: First, increase humans' awareness of the existence and risks of their own biases, and promote their initiation in combating these biases. Second, help humans to establish a physical and mental condition that is less vulnerable to biases. To meet the first goal, earlier literature in psychology has proposed various educational and training interventions to enhance humans's ability in debiasing in their future decision making [20, 75, 51, 85, 35, 90]. In contrast, much less attention has been paid with respect to the second goal. (2) *During-decision*: The main goal for designing during-decision intervention is to nudge humans to consciously adopt Type 2 processing. While such interventions could be as simple as forcing humans to slow down their decision process [12, 77], another promising but under-explored direction is to assist humans to formalize their thinking process (e.g., as a checklist of actions or if-then rules) and ground their decisions on sound data [84, 66]. (3) *Post-decision*: Finally, post-decision interventions can be designed to help humans to reflect upon and critique their decisions. These interventions aim to both enable humans to identify any potential biases that they have been subject to in their decisions, and allow humans to re-examine their decisions comprehensively and systematically.

Experiments and expected outcome. We plan to conduct randomized controlled experiments to examine how different interventions affect human decision making with machines in the loop. The overall experiment design would follow the design in Task 1.1 with the **independent variables** being the interventions. The **performance measure** will be whether the empirical prediction accuracy of Bayesian rationality becomes relatively higher compared with no interventions. This task will provides us an understanding of the effectiveness of different interventions in nudging humans towards being more Bayesian rational, which could play an important role when we aim to promote humans in making sensible decisions.

3.2.4 Task 2.3: Develop a machine learning framework for human behavioral modeling

In the above proposed tasks, we aim to understand human behavior under different scenarios and examine which behavioral models better explain human behavior. As a commonly practice in the literature, we measure how well a model explains real-world human behavior by examining its empirical prediction accuracy, i.e., how well the model *fit* the data. However, the goal of a behavioral model should be to explain and predict human behavior for even the unseen scenarios. This aligns well with how the field and problem setup of machine learning is structured. In addition, traditionally, these behavioral experiments were often conducted in the lab setting with only dozens of data points. Therefore, researchers are limited to test a small number of hypothesis to obtain statistically significant results, and therefore it might not benefit much by applying machine learning framework. In this proposed research, we aim to conduct larger scale experiments with more data points which would enable us to apply techniques and insights from machine learning to examine behavioral models.

Illustrative example. Consider the belief updating example. Again, let $P(e)$ denote the prior of event e and I denote the signal/information. If humans are Bayesian, they update their beliefs following Bayes rule $P(e|I) = \frac{P(I|e)P(e)}{P(I)}$. If they follow our alternative model, they update their beliefs following the rule $P(e|I) \propto P(I|e)^{\alpha} P(e)^{\beta} P(I)^{\gamma}$. To formulate this in machine learning terms, the prior $P(e)$ and signal distribution ($P(I|e)$ and $P(I)$) are the features that we can observe, and the posterior $P(e|I)$ are the labels we aim to predict. In our experiments, we collect data points by designing different features and observe humans responses as labels.

When we are trying to fit the data to examine whether humans are Bayesian, note that since there is no parameters to *tune* in the Bayes rule, there exists only a single hypothesis to fit, and we are able to bound the generalized prediction accuracy from empirical prediction accuracy using standard concentration bound. However, when we are trying to fit the data to our model, as in Equation (3), there are three variables (α,β, and γ) that we can tune to fit the data, which implies the power of the hypothesis set (which can be characterized by notions similar to VC dimension [101]) is larger and good empirical prediction accuracy might not mean good generalized prediction accuracy. In this task, we propose to utilize the machine learning theory to formally characterize this approximation-generalization trade-off and answer other questions.

Research questions. We plan to apply the framework of machine learning theory to answer and address the following questions in modeling human with behavioral experiments. (1) Quantify the expressiveness of different behavioral models) and refine the performance measure of the behavioral experiments. We plan to approach this problem by using VC-dimension style notions to quantify the expressiveness of the model. This also enables us to understand the gap between empirical accuracy and generalized accuracy to refine the performance measure. (2) Use machine learning to guide the design of behavioral experiments. Given that our experiment design involves choosing the features of data points to get labels, this experiment design problem can be naturally cast as active learning problems. We will explore how active learning might help improve the efficiency of data collection. (3) Examine whether new models are needed. With the expressiveness notion and the number of data points, we can check the inductive biases to see whether the behavioral model has reached its prediction limit, and whether a new alternative model is needed. For example, in settings with huge amount of human behavioral data and interpretability is less a concern, we might introduce neural network as a candidate human behavior model.

3.3 Thrust 3: Aligning Humans and Machines by Including Humans in the Loop

We demonstrate that the information design problem in machine-in-the-loop decision making can be formulated as a constrained optimization problem. However, we have abstracted away an important perspective of the problem formulation: what should the objective of the optimization be. While the goal of machine-in-the-loop decision making is to assist humans in making decisions, an ill-defined objective could pose potential concerns and lead to undesired outcomes. This problem is further amplified in high-stake domains when there are often multiple objectives to balance during decision making. Take homelessness prevention for example, homelessness service providers might have multiple objectives in mind when deciding how to allocate the resources (e.g., transitional housing, emergency shelters, or rapid re-housing) to families in need. They may want to (1) reduce the number of families that are not offered a shelter place, and (2) reduce the expected number of families that would re-enter the system again in the future. Since these objectives do not always align, homelessness service providers often need to trade-off and balance these objectives during decision making. How we should design assistive ML algorithms with the appropriate objective that aligns with human values?

In this research thrust, we propose to address this concern by including human decision-makers in the loop to help determine what the objective of optimization should be. We will first conduct the surveys to elicit the preferences over objectives from individual human decision-makers. In particular, we will elicit individual preferences with both the normative approach, in which decision-makers are asked to report the criteria of their decisions, and the descriptive approach, in which we ask decision-makers to make decisions in a set of hypothetical scenarios and infer their underlying criteria from their decisions. After eliciting individual preferences, we will explore methods in aggregating individual preferences into collective objectives. The proposed activities in this thrust will be based on collaborations with domain experts in real-world problems. For data collection, during the initial development phase, we will collect data using Amazon Mechanical Turk as proof of concepts. Later we will work with domains experts and evaluate our approaches in the real world. In particular, we will work with Prof. Patrick Fowler (see letter of collaboration) at WashU Brown School of Social Work on the problem of homelessness prevention. We will also leverage the interdisciplinary effort at WashU to apply the proposed research to domain problems in social sciences and healthcare.

3.3.1 Task 3.1: Elicit individual preferences: normative vs. descriptive approaches

In order to design objectives that align with human values, we need to elicit humans preferences. There are usually two lines of approaches in the literature. The first one is the normative approach, in which humans are asked to report their criteria of decision making, and the other one is the descriptive approach, in which the criteria is inferred from decisions made by humans. In this task, we aim to examine both approaches and also investigate whether iteratively apply both approaches can support deliberation and lead to more consistent reporting.

Experiments. In this task, we will start the investigation by conducting the experiments on Amazon Mechanical Turk. After giving tutorials of the task background (e.g., homelessness prevention), workers are going to answers questions on what their objective is for the task if they are the decision makers. There are two types of surveys. In survey A, workers will be given a set of criteria generated by domain experts. They will report how their own criteria aligns with the provided ones, in the forms of ranking, hierarchical order, or weighted sum. In survey B, workers will be given a set of hypothetical scenarios and need to provide their decisions in the given scenarios. We will then apply ML approaches to learn the underlying criteria (again, in the form of ranking, hierarchical order, or weighted sum). We will control the **independent variables** to be the amount of tutorials given (representing novice or expert decision-makers), whether workers answer survey A, B, or both, and different sets of expert-provided criteria. We will measure the results using distributions of answers (in survey A) and machine-learned criteria (in survey B). We will also perform additional qualitative survey questions to gauge users' decision process and obtain feedback in the survey design.

Expected outcome. The results will help us understand both on humans' criteria during decision making in given tasks and on how different factors (e.g., familiarity of the task, elicitation methods) impact the outcome of the elicitation. The results also provide insights on human decision process and shed lights on how to improve the design in the field.

3.3.2 Task 3.2: Aggregate individual preferences into collective objectives

With the individual preferences elicited, our next task is to aggregate these preferences into an objective for our assistive ML algorithm. We plan to explore two different approaches for aggregation under a range of settings and examine the outcome of aggregation through conducting surveys with both lay-persons and domain experts.

Proposed approaches. We will explore the aggregation with two different approaches. In the first one, we will leverage social choice theory, which is the main theory discussing how to aggregate multiple individual preferences that satisfy certain properties/axioms (one typical usage of the theory is in designing voting mechanisms), for aggregation. Given that it is often impossible to simultaneously satisfy all the axioms, different social choice rules are designed to satisfy a subset of axioms or approximately satisfy them. In the second approach, we plan to formulate this as a machine learning problem, treating each individual preference as data points, defining corresponding loss functions, and finding an objective that minimize the loss. Note that there are some corresponding between the two approaches, for example, the utilitarian rule (also called the max-sum rule) in social choice essentially corresponds to the squared loss function in the ML approach. However, there do not exist a obvious one-to-one mapping.

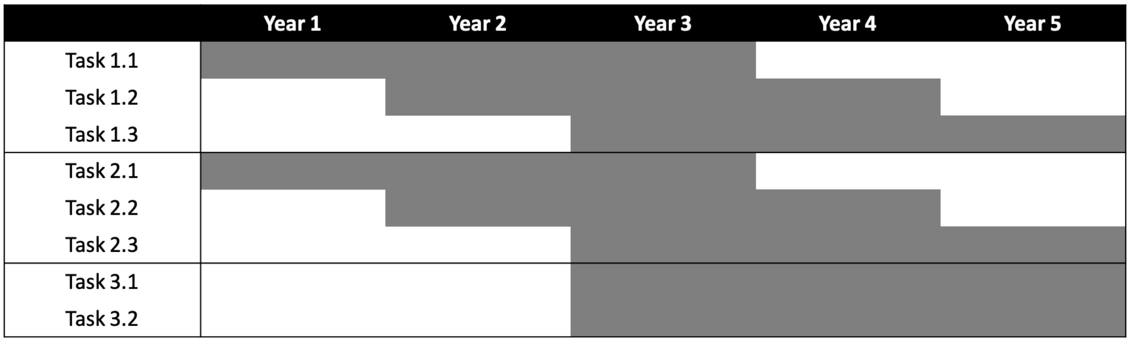
We will examine the aggregation outcome in two dimensions. First, we will examine the aggregated outcome through through the lens of the two approaches.i.e., what axioms are satisfied for ML aggregation and what loss is incurred for aggregation by social choice rules. In addition, we will recruit both domain experts and lay-persons (from Mechanical Turk) to obtain insights on which aggregation aligns more with what humans would do and also investigate underlying reasoning. The observations could in turn help us define axioms that should be satisfied during aggregation for the target application.

3.3.3 Task 3.3: Work with domain experts in real-world applications

We will work with domain experts for real-world applications. In particular, we plan to work with Prof. Patrick Fowler at WashU Brown School of Social Work on applying the machine-in-the-loop decision making framework for homelessness prevention. The proposed activities will be built on the PI's existing collaborations [29] with Prof. Patrick Fowler on designing online algorithms for allocating resources to homelessness households, where we explore algorithmic solutions for homelessness prevention with the objective of minimizing the expected number of families that might re-enter the system after obtaining the resources. In this task, we plan to survey domain experts and/or homelessness service providers to get a better insights of their decision-making process, their objectives in decision making, and the type of decision support that need to inform the type of information ML algorithms should provide. In addition to homelessness prevention, We will also leverage the interdisciplinary effort at WashU, including Division of Computational and Data Science (DCDS) and Center for Collaborative Human-AI Learning and Operation (HALO) to apply the work in the research to domain problems in social sciences and healthcare.

3.4 Evaluation Plan

The proposed research will span five years with the timeline as below. The tasks in the first two thrusts have been organized in a way that we plan to perform the tasks in a sequential manner. We will perform the tasks in Thrust 3, which address the negative impacts, after we have initial results for the first two thrusts.



	Year 1	Year 2	Year 3	Year 4	Year 5
Task 1.1	■	■	■		
Task 1.2		■	■	■	
Task 1.3			■	■	■
Task 2.1	■	■	■		
Task 2.2		■	■	■	
Task 2.3			■	■	■
Task 3.1			■	■	■
Task 3.2			■	■	■

For the evaluation of the proposed research, there are three main components:

- Data collection: The collected data of the behavioral experiments will be made publicly available to the research community. We believe the larger-scale behavioral data would be of important research value.
- Modeling: We will examine the proposed user behavior model through empirical prediction accuracy on the collected data. One of our research task (Task 1.3) involves developing more sensible performance measure than empirical accuracy.
- Theory: we will derive the performance guarantees (regret bounds or convergence rate) and analyze the computational complexity of the proposed learning and optimization algorithms. We will perform equilibrium analysis to characterize the human behavior in the equilibrium structure. Simulation will also be performed to evaluate the algorithm performance under the conditions both when users follow our proposed models and when users do not exactly follow to test for robustness of our proposed algorithms.

4 Education Plan

The PI aims to broaden the research participation and develop education plans that integrate with the proposed research throughout the duration of the CAREER project. To maximize the impacts of the proposed activities, the PI will collaborate with several existing programs at WashU.

4.1 Broadening Research Participation

This project will invest efforts in broadening the participation in computing, including developing activities to expose high-school students in research, actively recruiting female and underrepresented minority students, and engaging undergraduate research participation.

Activities for high-school students. The PI will partner with the Institute for School Partnership (ISP) at WashU to design activities for high-school students. The goal is to cultivate next-generation scientists and engineers through exposing high-school students to academic research and stimulating their interests in computing. We have allocated budgets for the planned activities (with the collaboration letter attached). In particular, the McKelvey School of Engineering at WashU has conducted a pilot camp in Summer 2021 for local high school students of low-income backgrounds, with administrative and educational support provided by ISP. This pilot is planned to become an annual summer workshop. The PI plans to develop a three half-days summer workshop "Human-Centered Machine Learning" within this framework. The workshop will consists of two main components. The first component will provide students a broad overview of machine learning (ML) and human behavior to get students familiar with how ML works and understand how humans often respond to ML systems. The second component consists of conducting group research projects guided by Ph.D. students. We will prepare data sets and existing ML modules. Students will explore different system designs (grounded by research activities in this proposal) for ML to assist human decision-making and investigate the benefits and pitfalls of each design. In the first two summers, we will work with ISP and recruit a local high-school teacher during the summer to help develop the workshop. The teachers will get exposed to ongoing research in the field and work with the PI in identifying topics that will better motivate and engage high-school students. In the third summer, we will host a workshop with around 25 high-school teachers to disseminate the course design to maximize the potential outreach and obtain additional feedback. We will then host the workshop in year 4 and 5 by recruiting high-school students from schools that work with ISP.

Evaluation plan: The ISP will help with the logistics of the workshop, including recruiting high-school teachers and students, and also help with evaluations after the launch. In particular, we will conduct anonymous survey, with the help and consultation from ISP, to students before and after the workshop to evaluate their understanding of the topic and their aspirations in pursuing higher-education in STEM.

Engagements of female and underrepresented minority students. The PI will actively recruit female students for joining the research. The PI has worked with three female undergraduate students (out of nine undergraduate students that worked with the PI) at Washington University. Two of them have continued their Ph.D. studies after graduation (at Stanford and Duke) and one of them has been going to the industry (at Google). Washington University is actively committed to the goal of increasing the representation of women at the Ph.D. level. For example, the CSE department, the McKelvey School of Engineering, and the Provost's Office of Diversity together fund a Platinum Sponsorship of Grace Hopper.

The PI plans to leverage the effort of WashU to offer research opportunities to under-represented students. The PI has currently been advising one underrepresented undergraduate student during the summer of 2021 through WashU Summer Engineering Fellowship (WUSEF), which provides funds for students from backgrounds underrepresented in the STEM fields to perform summer research. In addition to working with WUSEF each summer, the PI will also seek collaboration opportunities with the AI4ALL consortium that

WashU will be joining with a focus on AI education programs, targeting underrepresented high school students, and the Missouri Louis Stokes Alliance for Minority Participation (MOLSAMP), of which WashU is one of the participating institutions, for offering summer research opportunities for minority participation.

Undergraduate research participation. Undergraduate students will be heavily engaged in the proposed research. The PI has been actively involved in the NSF REU site "Big Data Analytics" at WashU, and the results have led to a publication [64] with undergraduate students at the ACM Conference on Economics and Computation (EC), one of the top and most selective venues at the interface of economics and computations. The PI is also currently working with the WashU Summer Engineering Fellowship Program (WUSEF) and is advising an under-represented undergraduate student. The students the PI advised at the REU site have all continued their Ph.D. studies in the Computer Science field (at UT Austin, Duke, and CMU) after graduation. The PI is committed to annually support REU/WUSEF research projects inspired by this proposal, such as understanding user behavior in computational systems through conducting behavioral experiments or analyzing existing datasets. The PI will also support undergraduate students on independent research projects during the academic year.

4.2 Course and Teaching Development

The research goal of the PI is to combine the strengths of both humans and machine learning (ML) to solve tasks neither can solve alone. To achieve this goal, we need to advance our understanding of ML, humans, and the interactions between them. Correspondingly, the education goal of the PI is to prepare students in these fronts. To achieve this education goal, the PI has been regularly teaching two courses: undergraduate-level *CSE 417T: Introduction to Machine Learning* and graduate-level *CSE 518A: Human-in-the-Loop Computation*. As part of this CAREER project, the PI plans to heavily revise *CSE 518A* into a new course *Human-AI Interaction and Collaboration*. In addition to the general coverage of ML and human modeling (from behavioral economics, psychology, and HCI), there will be two main themes for the course topics. First. we will cover and discuss human-in-the-loop machine learning, addressing the techniques of incorporating humans in the learning process to advance machine learning. Second, we will discuss topics with a human-centered focus, including how humans process information from ML (such as interpretability, trustworthiness, and topics explores in this proposal) and how ML impacts human welfare (such as fairness, privacy, and ethical concerns). We will also include practical domain applications in social sciences and healthcare in the course materials (in the form of assignments, projects, or guest lectures) by leveraging the Division of Computational and Data Science (DCDS) and the Center for Collaborative Human-AI Learning and Operation (HALO) at WashU. The course materials will be made public available online to enable self-study or to be used in other institutions.

The PI will deploy active learning techniques, such as peer instruction [21], in delivering the course. In fact, the research activities in this proposal have implications on how we shape our teaching methods in education, in which the goal of the instructor (mapped to machine learning in the proposed research) aims to communicate with the students (mapped to human decision makers) to enable better learning. For example, my prior works [44, 93] showed that by providing peer information (the feedback from others) to crowd workers can help improve their short-term work performance. In addition, by coupling the peer information with expert information, we could improve the long-term work performance of crowd workers. The results not only align with the observations of peer instruction [21] in education that students achieve better learning when being able to discuss with the peers, but also provide a potential algorithmic framework on determining when and how to enable peer instruction. The PI plans to leverage this synergy to both inspire research questions from teaching practices and improve education from research insights.

Evaluation plan. [CJ: to be updated] The PI will work with the Center for Integrative Research on Cognition, Learning, and Education (CIRCLE) at WashU to develop evaluation plan for the proposed course. The evaluations will be conducted based multiple metrics, including whether students obtain firm grasp of the subject and on whether the course motivates students in applying the knowledge in different domains.

The PI will coordinate with the Teaching Center at WashU to periodically videotape and assess the lectures from the course. The PI will also attend the events and workshops organized by the Teaching Center to improve his skills as an educator.

5 Broader Impacts

This research has a direct impact on the design for a broad range of online platforms, including recommendation systems, user-generated content platforms, systems, social-networking sites, and other platforms with active human participation. In addition, as algorithmic decision making has been deployed more widely in policy making, this research also contributes to improve decision making for societal issues. In particular, the PI has existing collaborations with Prof. Patrick Fowler at Brown School of Social Work to study the allocation of scarce resources for homeless prevention [29] and with Dr. Jason Wellen at Medical School to apply computational approaches for living donor kidney transplantation [64]. The PI plans to continue and expand the collaborations through the Division for Computational and Data Sciences (DCDS) which brings together the department of Computer Science & Engineering with the departments of Political Science and Psychological and Brain Sciences in Arts & Sciences and with the Brown School of Social Works, to apply the machine-in-the-loop decision-making framework to address societal issues. Moreover, the PI is the member of the newly found Center for Collaborative Human-AI Learning and Operation (HALO) at Washington University, which provides additional collaboration opportunities with the medical school at Washington University on healthcare problems.

Dissemination of results. One of the main research effort in this proposed research is to collect human behavioral data through multiple sets of large-scale behavioral experiments. We plan to make the collected publicly accessibly to the research community. To disseminate our research results to a broad audience, in addition to regular conference and journal publications, we will publicly release the software implementations of algorithms, simulation test-bed, and models developed in this project. Furthermore, we will disseminate results within the interdisciplinary DCDS program at Washington University through regular interaction with other faculty in the program, as well as its seminar series.

6 Results from Prior NSF Support.

Dr. Ho is the co-PI on the current grant ("FAI: FairGame: An Audit-Driven Game Theoretic Framework for Development and Certification of Fair AI", IIS 1939677, $444,145, Jan 2020 to Dec 2022). *Intellectual Merit*: This project provides a general game theoretical framework for fair decision making and auditing in stochastic, dynamic environments. In addition, we develop a new foundational understanding of the legal landscape pertaining to fair decisions, particularly as applied in the context of the proposed framework. The project so far has generated four publications (Estornell et al., 2021; Nguyen et al., 2021; Raderv et al., 2020; Tang et al., 2021). *Broader Impacts*: The work is supporting the training of graduate students and the development of new auditing algorithms that have impacts to AI and society. New curriculum development efforts, including a new course on AI and Society, as well as active participation in the interdisciplinary computational and data science program at Washington University, have helped disseminate the research to a broad and diverse audience.

References

[1] Jacob Abernethy, Yiling Chen, Chien-Ju Ho, and Bo Waggoner. Low-cost learning via active data procurement. In *16th ACM Conf. on Economics and Computation (EC)*, 2015.

[2] Tal Alon, Magdalen Dobson, Ariel Procaccia, Inbal Talgam-Cohen, and Jamie Tucker-Foltz. Multiagent evaluation mechanisms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[3] Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256, 2002. Preliminary version in *15th ICML*, 1998.

[4] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018.

[5] Kay W Axhausen and Tommy Gärling. Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems. *Transport reviews*, 12(4):323–341, 1992.

[6] Ashwinkumar Badanidiyuru, Kshipra Bhawalkar, and Haifeng Xu. Targeting and signaling in ad auctions. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2545–2563. SIAM, 2018.

[7] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 2019.

[8] Dirk Bergemann and Stephen Morris. Information design: A unified perspective. *Journal of Economic Literature*, 57(1):44–95, 2019.

[9] Michael Brückner and Tobias Scheffer. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 547–555, 2011.

[10] Michael Brückner, Christian Kanzow, and Tobias Scheffer. Static prediction games for adversarial learning problems. *Journal of Machine Learning Research*, 13(Sep):2617–2654, 2012.

[11] Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012.

[12] E Moulton Carol-anne, Glenn Regehr, Maria Mylopoulos, and Helen M MacRae. Slowing down when you should: a new model of expert judgment. *Academic Medicine*, 82(10):S109–S116, 2007.

[13] Gabriel Carroll. Robustness and linear contracts. *American Economic Review*, 105(2):536–63, 2015.

[14] Gabriel Carroll. Robustness and separation in multidimensional screening. *Econometrica*, 85(2): 453–488, 2017.

[15] Gabriel Carroll and Ilya Segal. Robustly optimal auctions with unknown resale opportunities. *The Review of Economic Studies*, 86(4):1527–1555, 2019.

[16] Sylvain Chassang. Calibrated incentive contracts. *Econometrica*, 81(5):1935–1971, 2013.

[17] Yiling Chen, Chara Podimata, Ariel D. Procaccia, and Nisarg Shah. Strategyproof linear regression in high dimensions. In *Proceedings of the 2018 ACM Conference on Economics and Computation (EC)*, 2018.

[18] Sharath R. Cholleti, Sally A. Goldman, Avrim Blum, David G. Politte, and Steven Don. Veritas: Combining expert opinions without labeled data. In *Proceedings 20th IEEE international Conference on Tools with Artificial intelligence (ICTAI)*, 2008.

[19] Vincent Conitzer, Markus Brill, and Rupert Freeman. Crowdsourcing societal tradeoffs. In *AAMAS*, pages 1213–1217, 2015.

[20] Pat Croskerry, Geeta Singhal, and Sílvia Mamede. Cognitive debiasing 1: origins of bias and theory of debiasing. *BMJ quality & safety*, 22(Suppl 2):ii58–ii64, 2013.

[21] Catherine Crouch and Eric Mazur. Peer instruction: Ten years of experience and results. *Am. J. Phys.*, 69(9):970–977, September 2001.

[22] Tianjiao Dai and Juuso Toikka. Robust incentives for teams. *Unpublished manuscript, Mass. Inst. of Technology, Cambridge, MA*, 2017.

[23] A. P. Dawid and A. M. Skene. Maximum likeihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, 28:20–28, 1979.

[24] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38, 1977.

[25] Peter Diamond. Managerial incentives: on the near linearity of optimal compensation. *Journal of Political Economy*, 106(5):931–957, 1998.

[26] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114, 2015.

[27] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3): 1155–1170, 2018.

[28] Bolin Ding, Yiding Feng, Chien-Ju Ho, and Wei Tang. Competitive information disclosure with multiple receivers. Working paper, 2021.

[29] Zehao Dong, Sanmay Das, Patrick Fowler, and Chien-Ju Ho. Efficient nonmyopic online allocation of scarce resources. In *Proceedings of the 20th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2021.

[30] Xiaoni Duan, Chien-Ju Ho, , and Ming Yin. Do diverse interactions mitigate biases in crowdwork? an experimental study. In *The 8th AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2020.

[31] Shaddin Dughmi and Haifeng Xu. Algorithmic bayesian persuasion. *SIAM Journal on Computing*, 2019.

[32] Mary T Dzindolet, Linda G Pierce, Hall P Beck, and Lloyd A Dawe. The perceived utility of human and automated aids in a visual detection task. *Human Factors*, 44(1):79–94, 2002.

[33] Yuval Emek, Michal Feldman, Iftah Gamzu, Renato PaesLeme, and Moshe Tennenholtz. Signaling schemes for revenue maximization. *ACM Transactions on Economics and Computation (TEAC)*, 2 (2):1–19, 2014.

[34] Jonathan St BT Evans and Keith Ed Frankish. *In two minds: Dual processes and beyond.* Oxford University Press, 2009.

[35] Rebecca Jean Featherston, Aron Shlonsky, Courtney Lewis, My-Linh Luong, Laura E Downie, Adam P Vogel, Catherine Granger, Bridget Hamilton, and Karyn Galvin. Interventions to mitigate bias in social work decision-making: A systematic review. *Research on Social Work Practice*, 29(7): 741–752, 2019.

[36] Noah D Goodman, Joshua B. Tenenbaum, and The ProbMods Contributors. Probabilistic Models of Cognition. `http://probmods.org/v2`, 2016. Accessed: 2021-6-19.

[37] Ben Green and Yiling Chen. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 90–99, 2019.

[38] 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.

[39] Thomas L. Griffiths and Joshua B. Tenenbaum. Optimal predictions in everyday cognition. *Psychological Science*, 17(9):767–773, 2006.

[40] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42, 2018.

[41] Lars Peter Hansen and Thomas J Sargent. Three types of ambiguity. *Journal of Monetary Economics*, 59(5):422–445, 2012.

[42] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pages 111–122, 2016.

[43] Chien-Ju Ho and Jennifer Wortman Vaughan. Online task assignment in crowdsourcing markets. In *26th AAAI Conference on Artificial Intelligence (AAAI)*, 2012.

[44] Chien-Ju Ho and Ming Yin. Working in pairs: Understanding the effects of peer communication in crowdwork, 2018. Working Paper.

[45] Chien-Ju Ho, Yu Zhang, Jennifer Wortman Vaughan, and Mihaela van der Schaar. Towards social norm design for crowdsourcing markets. In *4th Human Computation Workshop (HCOMP)*, 2012.

[46] Chien-Ju Ho, Shahin Jabbari, and Jennifer Wortman Vaughan. Adaptive task assignment for crowdsourced classification. In *30th Intl. Conf. on Machine Learning (ICML)*, 2013.

[47] Chien-Ju Ho, Aleksandrs Slivkins, and Jennifer Wortman Vaughan. Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. In *15th ACM Conf. on Electronic Commerce (EC)*, 2014.

[48] Chien-Ju Ho, Aleksanrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. Incentivizing high quality crowdwork. In *24th Intl. World Wide Web Conf. (WWW)*, 2015.

[49] Chien-Ju Ho, Rafael Frongillo, and Yiling Chen. Eliciting categorical data for optimal aggregation. In *30th Advances in Neural Information Processing Systems (NIPS)*, 2016.

[50] Chien-Ju Ho, Aleksandrs Slivkins, and Jennifer Wortman Vaughan. Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. *Journal of Artificial Intelligence Research*, 55:317 – 359, 2016.

[51] Milos Jenicek. *Medical error and harm: Understanding, prevention, and control*. CRC Press, 2010.

[52] Rong Jin and Zoubin Ghahramani. Learning with multiple labels. In *Advances in Neural Information Processing Systems (NIPS)*, 2003.

[53] D. Kahneman and A. Tversky. On the psychology of prediction. *Psychological Review*, 80:237–251, 1973.

[54] Daniel Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

[55] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decisions under risk. *Econometrica*, pages 263–291, 1979.

[56] Emir Kamenica. Bayesian persuasion and information design. *Annual Review of Economics*, 11: 249–272, 2019.

[57] Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6): 2590–2615, 2011.

[58] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.

[59] Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 825–844, 2019.

[60] Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocations rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

[61] 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.

[62] Vivian Lai, Han Liu, and Chenhao Tan. " why is' chicago'deceptive?" towards building model-driven tutorials for humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[63] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, et al. Webuildai: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–35, 2019.

[64] Zhuoshu Li, Kelsey Lieberman, William Macke, Sofia Carrillo, Chien-Ju Ho, Jason Wellen, and Sanmay Das. Incorporating compatible pairs in kidney exchange: A dynamic weighted matching model. In *Proceedings of the 2019 ACM Conference on Economics and Computation (EC)*, 2019.

[65] Yang Liu and Chien-Ju Ho. Incentivizing high quality user contributions: New arm generation in bandit learning. In *32nd AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[66] Joseph J Lockhart and Saty Satya-Murti. Diagnosing crime and diagnosing disease: bias reduction strategies in the forensic and clinical sciences. *Journal of forensic sciences*, 62(6):1534–1541, 2017.

[67] George Loewenstein. Out of control: Visceral influences on behavior. *Organizational behavior and human decision processes*, 65(3):272–292, 1996.

[68] Jennifer M Logg, Julia A Minson, and Don A Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103, 2019.

[69] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 2017.

[70] George J. Mailath and Larry Samuelson. Learning under diverse world views: Model-based inference. *American Economic Review*, 110(5):1464–1501, May 2020. doi: 10.1257/aer.20190080. URL `https://www.aeaweb.org/articles?id=10.1257/aer.20190080`.

[71] Daniel McFadden. Econometric models of probabilistic choice. *Structural analysis of discrete data with econometric applications*, 198272, 1981.

[72] Daniel McFadden. Economic choices. *American economic review*, 91(3):351–378, 2001.

[73] Jianjun Miao and Alejandro Rivera. Robust contracts in continuous time. *Econometrica*, 84(4): 1405–1440, 2016.

[74] Stephen Morris. The common prior assumption in economic theory. *Economics & Philosophy*, 11 (2):227–253, 1995.

[75] Tess Neal and Stanley L Brodsky. Forensic psychologists' perceptions of bias and potential correction strategies in forensic mental health evaluations. *Psychology, Public Policy, and Law*, 22(1):58, 2016.

[76] Ritesh Noothigattu, Snehalkumar Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel Procaccia. A voting-based system for ethical decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

[77] Eoin D O'Sullivan and Susie J Schofield. A cognitive forcing tool to mitigate cognitive bias–a randomised control trial. *BMC medical education*, 19(1):1–8, 2019.

[78] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–52, 2021.

[79] Drazen Prelec. The probability weighting function. *Econometrica*, pages 497–527, 1998.

[80] Gonzalo Ramos, Jina Suh, Soroush Ghorashi, Christopher Meek, Richard Banks, Saleema Amershi, Rebecca Fiebrink, Alison Smith-Renner, and Gagan Bansal. Emerging perspectives in human-centered machine learning. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2019.

[81] Vikas Raykar, Shipeng Yu, Linda Zhao, Gerardo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322, 2010.

[82] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[83] Marc Oliver Rieger and Mei Wang. Cumulative prospect theory and the st. petersburg paradox. *Economic Theory*, 28(3):665–679, 2006.

[84] Dennis Rosen. The checklist manifesto: How to get things right. *JAMA*, 303(7):670–673, 2010.

[85] Anne-Laure Sellier, Irene Scopelliti, and Carey K Morewedge. Debiasing training improves decision making in the field. *Psychological science*, 30(9):1371–1379, 2019.

[86] Rajiv Sethi and Muhamet Yildiz. Communication with unknown perspectives. *Econometrica*, 84(6): 2029–2069, 2016.

[87] Kenneth A Small. A discrete choice model for ordered alternatives. *Econometrica: Journal of the Econometric Society*, pages 409–424, 1987.

[88] Keith E Stanovich and Richard F West. On the relative independence of thinking biases and cognitive ability. *Journal of personality and social psychology*, 94(4):672, 2008.

[89] Ola Svenson. Process descriptions of decision making. *Organizational behavior and human performance*, 23(1):86–112, 1979.

[90] Carl Symborski, Meg Barton, Mary Quinn, C Morewedge, K Kassam, James H Korris, and CA Hollywood. Missing: A serious game for the mitigation of cognitive biases. In *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*, pages 1–13, 2014.

[91] Wei Tang and Chien-Ju Ho. Bandit learning with biased human feedback. In *Proceedings of the 18th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2019.

[92] Wei Tang and Chien-Ju Ho. On the bayesian rationality assumption in information design. Working paper, 2021.

[93] Wei Tang, Chien-Ju Ho, and Ming and Yin. Leveraging peer communication to enhance crowdsourcing. In *The World Wide Web Conference (WWW)*, 2019.

[94] Wei Tang, Chien-Ju Ho, and Yang Liu. Bandit learning with delayed impact of actions. Working paper, 2020.

[95] Wei Tang, Chien-Ju Ho, and Yang Liu. Differentially private contextual dynamic pricing. In *Proceedings of the 19th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*, 2020.

[96] Wei Tang, Chien-Ju Ho, and Yang Liu. Optimal query complexity of secure stochastic convex optimization. In *34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[97] Wei Tang, Chien-Ju Ho, and Yang Liu. Linear models are robust optimal under strategic behavior. In *24nd Intl. Conf. on Artificial Intelligence and Statistics (AISTATS)*, 2021.

[98] Joshua Tenenbaum. Bayesian modeling of human concept learning. In *Advances in Neural Information Processing Systems*. MIT Press, 1999.

[99] Kenneth E Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.

[100] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974. doi: 10.1126/science.185.4157.1124.

[101] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

[102] John von Neumann and Oscar Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, 1944.

[103] Jacob Whitehill, Ting fan Wu, Jacob Bergsma, Javier R. Movellan, and Paul L. Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[104] Timothy D Wilson and Nancy Brekke. Mental contamination and mental correction: unwanted influences on judgments and evaluations. *Psychological bulletin*, 116(1):117, 1994.

[105] George Wu and Richard Gonzalez. Curvature of the probability weighting function. *Management science*, 42(12):1676–1690, 1996.

[106] Doris Xin, Litian Ma, Jialin Liu, Stephen Macke, Shuchen Song, and Aditya Parameswaran. Accelerating human-in-the-loop machine learning: challenges and opportunities. In *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning*, pages 1–4, 2018.

[107] Michael Yeomans, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg. Making sense of recommendations. *Journal of Behavioral Decision Making*, 32(4):403–414, 2019.

[108] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12, 2019.

[109] Fabio Massimo Zanzotto. Human-in-the-loop artificial intelligence. *Journal of Artificial Intelligence Research*, 64:243–252, 2019.

[110] Yudian Zheng, Guoliang Li, Yuanbing Li, Caihua Shan, and Reynold Cheng. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment*, 10(5):541–552, 2017.