Behavior-Informed AI: Factoring in Human Behavior for Robust Learning and Improved Decision-Making Assistance

Chien-Ju Ho, Washington University in St. Louis

Overview

The rapid advancement of artificial intelligence (AI) has revolutionized the way we approach problem-solving, innovation, and interaction with technology. Central to this evolution is AI's ability to learn from vast amounts of human-generated and annotated data, empowering machines to mimic and even surpass human capabilities across various domains. As AI continues to progress, there lies immense potential to harness its power to augment and enhance human decision-making, ultimately fostering more informed choices and improved outcomes. However, humans are known to make imperfect or even biased decisions. This impedes AI development by introducing biases into the data. To optimally assist humans and prevent them from succumbing to these biases, it is crucial for machine learning algorithms to understand and incorporate knowledge of human behavior and biases.

In this proposal, our goal is to design behavior-informed AI that understands and accounts for human behavior in data generation and decision-making, leading to AI systems that are robust to biased training data and are able to enhance human decision making. To achieve this goal, we will pursue the following research threads:

- Understand and model human behavior: We plan to conduct behavioral experiments to examine human behavior in the context of data annotation and decision-making. Subsequently, we will develop interpretable and accurate human models by leveraging cognitive sciences and machine learning.
- Train AI to be robust to human biases: The framework involves curating bias-mitigated datasets through designing cognitive-grounded bias-mitigation interventions during data collection, as well as developing post-hoc learning algorithms that account for human biases during training.
- Create behavior-aware assistive AI: We will design assistive AI systems that account for human behavior and biases to enhance human decision making. The assistive AI will adaptively provide structured assistance and update the decision-making environment based on insights derived from human behavior.

Intellectual Merit

This proposed research will contribute to the empirical understanding of human behavior in the context of human-AI interactions. It will also provide theoretical foundations for studying the interactions between humans and AI algorithms, through incorporating human models in both learning frameworks and assistive AI design. The results of the proposal will provide insights into 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.

Broader Impacts

This research has a direct impact on the design of a broad range of online platforms with active human participation. Moreover, it also contributes to improve policy making for societal issues. In particular, the PI has existing collaborations with domain experts in the department of Psychology and Brain Sciences, Brown School of Social Work, and Medical School that apply computational approaches to practical problems such as allocation of scarce resources for homeless prevention and living donor kidney transplantation. The PI plans to continue and expand the collaborations through the Center for Collaborative Human-AI Learning and Operation (HALO) and the Division for Computational and Data Sciences (DCDS) at the Washington University in Saint Louis to apply the developed machine-in-the-loop decision making framework to address societal issues. This proposal also includes a comprehensive plan for enhancing the education and broadening the research outreach, including creating a course in human-AI collaborations, engaging undergraduate

research, and hosting summer workshops for high-school students.

Behavior-Informed AI: Factoring in Human Behavior for Robust Learning and Improved Decision-Making Assistance

1 Introduction

Machine learning (ML) has gained significant progress in the past decade. With the improving predictive power, ML has been increasingly involved in 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 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, humans are still naturally the final decision makers.

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 this decision-making paradigm presents significant promises, there are also challenges. In particular, humans are known to exhibit behavioral biases in making decisions. How do we take human behavior into account when designing ML algorithms to assist humans? Moreover, there might be multiple objectives to balance and trade-off during decision making, how should we decide the objective that aligns with human values?

In this research proposal, we 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 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 machine-in-the-loop decision making that accounts for different human models of decision making. In particular, we will explore the usage of information design, i.e., how ML should present information to humans such that humans can make better decisions. The information can be structured either as a recommendation of action, or a distribution of signals conditional on the state of the world. 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 human 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 decisions, 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.

- **Aligning the Objectives of Humans and Machines by Including Humans in the Loop**: In the above discussion, we formulate the information design in machine-in-the-loop decision making as an optimization problem. However, 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 objective could lead to undesired outcomes. In this thrust, we plan to borrow ideas from participatory design for algorithmic governance [59, 6, 72, 18] 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 this 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 machine-in-the-loop 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 studying the interactions between humans and ML, from the perspectives of machine learning, algorithmic economics, and online behavioral social science. From the machine learning perspective, the PI has explored the problem of learning from noisy human-generated data and optimally matching humans with suitable tasks [40, 43]. The PI also designed data elicitation mechanisms to enable more efficient learning [1, 46, 41]. From the economics perspective, the PI has explored the design of different types of incentives, such as reputation [42], monetary payments [44, 47], and attention [61]. Moreover, the PI has explored behavioral aspects of humans in computational environments [86, 92, 87] and address ethical considerations [90, 89]. 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 [45], how users are influenced by other users [88, 28], and how users make decisions under uncertainty [87]. The PI is active in the research communities, including organizing workshops at NIPS (now NeurIPS) and HCOMP to explore the connection of crowdsourcing and machine learning and to foster the study of theoretical foundations of human computation. The PI 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 area chair, senior program committee, and program committee in major AI/ML 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 in our everyday life, it is of paramount importance to study how machine learning 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 [104, 101], in which machine learning 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 recent years, including improving the overall performance of human-AI partnerships [35, 36, 57, 9, 58] and also in investigating the interpretability [79, 65, 38, 77, 53, 74] and trustworthiness [24, 30, 25, 64, 102, 103] 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 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 considers humans as data contributors. In this category, humans are often modeled as data sources that output data by drawing from a distribution according to a generative model. This assumption enables the research works in the literature on label aggregation and truth discovery to incorporate human data in machine learning [23, 78, 17, 48, 98, 22, 105, 16]. 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 [97, 12, 11, 39, 54, 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 human 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.

Belief updating. Bayesian models have been the prominent model for belief updating in algorithmic works [93, 37, 34]. However, it has been consistently observed in empirical studies that humans often deviate from being Bayesian [49, 95, 7, 85, 63, 68]. While there have been some alternative models in how humans update their beliefs [70, 82, 66, 100, 80, 75], they are not widely adopted in algorithmic frameworks. Formally, 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

$$P(e|I) \propto P(I|e)^{\alpha} P(e)^{\beta} P(I)^{\gamma} \tag{1}$$

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.

Decision making under uncertainty. The common assumption is expected utility theory [97] 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 [51]. Another commonly used theory, also Nobel-winning, is the discrete choice model [67, 83, 94], which accounts for the inherent randomness of human decision making by incorporating noises in the utility. Formally, 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. The above theories can be summarized below:

- Expected utility theory: it predicts that humans will take the action that maximizes $\sum_{k=1}^{K} p_k v(x_k)$.
- Prospect theory: 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: It predicts that humans will take the action that maximizes $\sum_{k=1}^{K} p_k v(x_k) + \epsilon$, where ϵ is the additional noise term that incorporates the intrinsic randomness of human 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 develop algorithmic frameworks in designing information from ML to assist humans, empirically examine human behavior in the context of ML-assisted decision making, and including humans in the loop to ensure the objectives of ML align with human values.

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

In this thrust, we plan to develop algorithmic frameworks for decision making with machines in the loop. As the main theme of this research proposal, we will explore the usage of *information design*, i.e., what types of information policy should ML choose to provide to human decision makers. As a simple illustrative example, consider a ML-assisted navigation system, in which ML provides information about traffic to humans, and humans can then decide what route to take based on both their own prior (previous experience and knowledge) and ML-provided information. The goal of ML is to optimize some pre-specified objective (e.g., minimizing the total transit time) while considering the human driver's response to the information (e.g., drivers might rely on their prior experience on the traffic more than the ML-provide information). The goal is to designing the information to present to humans[1]. The information can be presented in the form of a recommendation of action (e.g., which route to take conditional on the traffic) or a distribution of signals conditional on the world state (e.g., the estimated time for some particular routes given the traffic). This information design question is ubiquitous in machine-assisted decision making, e.g., how should ML provide information to online users to purchase the recommended products, to homelessness service providers to decide how to allocate resources, to medical doctors to decide what treatments to apply to patients?

To formulate this information design framework, we will start with the 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 (Task 1.1). We will then relax the full-knowledge assumption and investigate the associate learning (Task 1.2) and robust design problems (Task 1.3).

Prior work. The proposed activities in this thrust will be built on my prior works. In particular, the information design problem can be formulated as a *Stackelberg game*, in which ML first decides on the information policy, and humans decide on what decision to take based on the provided information. My prior works have explored Stackelberg games in a range of different domain applications. I have studied problems of contract design [47], in which the firm posts a contract and then the worker decides on the

[1] We note that our focus is on the design of *information structure*. While the presentations of information, such as visualization or language usages, are also important aspects of information design, they are not the focus of this proposed research.

amount of effort in response to the contract, learning with strategic responses [92], in which the learner posts a decision rule and then the agent responds with the goal of receiving a favorably treatment, and Bayesian persuasion [26, 87], in which the sender decides an information disclosure strategy to persuade the receiver to take certain actions. My most relevant work in the persuasion setting aligns well with this research if we consider the sender as ML and the receiver as human decision makers.

3.1.1 Task 1.1: Develop an optimization framework for information design

In this task, we aim to develop an information design framework that accounts for different human models.

Setting up the optimization framework. We consider the setting in which ML decides on an information policy and then the human decision maker takes action based on both her prior information and the ML-provided information. Below we use ML and human to denote the two roles respectively. We extend the classical information design framework by Kamenica and Gentzkow [52] (they only consider cases when humans are Bayesian rational) to account for different behavioral models. 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 policy ML chooses. Upon receiving the realization of the signal σ based on the information policy, human can choose an action a from an action set $\mathcal{A}$. To incorporate human behavioral models, we specify human behavior with two functions, a belief updating function $\omega(\mu_0, \sigma)$, which denotes the posterior distribution induced by the signal σ and prior μ_0, and a decision function $P(a|\omega(\mu_0, \sigma))$, characterized by a distribution of decisions given the posterior. The examples of the realization for the two functions can be found in Section 2.2. Let $V(a, \theta)$ be ML's utility when the human takes action a and the state of the world be θ. In this task, we assume $V(a, \theta)$ is given and known. We will address how to design this objective function in Thrust 3.[2] With these notations, the information design problem can be formulated as an optimization problem: $\max_\tau \ \mathbb{E}_{\theta,\sigma\sim\tau}[\sum_a P(a|\omega(\mu_0, \sigma))V(a, \theta)]$.

Following the seminal work by Kamenica and Gentzkow [52], without loss of generality, we can limit the space of information policy to be the distributions of posteriors, i.e., $\tau \in \Delta(\Delta(\Theta))$, with the constraint that the induced posterior need to be *plausible* (the posterior is possible to be induced by some information policy), i.e., $\tau \in \mathcal{K}$, where $\mathcal{K} = \{\tau \in \Delta(\Delta(\Theta)) : \exists\sigma \text{ such that } \forall\mu \in \mathsf{supp}(\tau), \mu = \omega(\mu_0, \sigma)\}$. Note that, when humans are Bayesian, this constraint reduces to the classical *Bayesian-plausibility*, i.e., $\mathbb{E}_{\mu\sim\tau}[\mu] = \mu_0$. Therefore, the information design problem can also be written as

$$\max_{\tau\in\Delta(\Delta(\Theta))} \ \mathbb{E}_{\theta,\mu\sim\tau}\left[\sum_{a\in\mathcal{A}} P(a|\mu)V(a, \theta)\right] \quad \text{s.t.} \quad \tau \in \mathcal{K}. \tag{2}$$

Research questions. With the optimization formulation in place, in this task, we aim to characterize and explore the information design problem with different human models. First, consider the classical case that humans are Bayesian rational, the decision function $P(a|\omega(\mu_0), \tau))$ 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 $\mu = \omega(\mu_0, \tau)$), the decision function is in the form of: $P(a|\mu) = \frac{\exp\left(\beta u^R(a|\mu)\right)}{\sum_{a'}\exp\left(\beta u^R(a'|\mu)\right)}$, where $u^R(a|\mu)$ is human's utility for taking action a with posterior μ. This is essentially a continuous softmax function. With this human behavioral model, the objective of the optimization problem is continuous differentiable, and first-order optimization techniques might be applied. The above discussion highlights the need to understand how different human behavior models impact the problem of information design for machine-in-the-loop decision making. In this task, we will assume the knowledge of human models and address the corresponding optimization problem.

[2] In some domains, such as recommendation system, the utility could be easier specified (e.g., the utility could be 1 when users follow the recommendation and 0 otherwise). However, in domains with more complex objectives, such as homelessness prevention, finding the right objective is challenging. We explore this issue more deeply in Thrust 3.

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 [29, 31, 8] 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 follows the discrete choice model or other models that lead to stochastic decision making, the optimization objective can usually be written as a continuous differentiable function. This enables the first-order optimization methods, such as gradient descent, to be applied. 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. The key element is to quantify the *smoothness* of human models, i.e., how much human behavior changes with a small change of provided information. My prior work on the convergence rate of secure convex optimization [91] will serve as the technical foundation for this problem.

3.1.2 Task 1.2: Design information with uncertain human behavior: A bandit approach

We start our investigation by assuming full knowledge of human behavior. While this assumption could be approximately satisfied when we have access to an abundant amount of human behavior data, it is generally a strong assumption that might not hold in practice. In this task, we relax this assumption and consider the scenario in which ML can repeatedly interact with humans over time. This scenario provides the opportunity for ML to infer humans' behavior models by utilizing the interaction of previous rounds (e.g., by observing past human decisions) and update the information policy (information scheme) in the future rounds.

Research questions. This problem naturally leads to the classical trade-off between exploration (choosing policy with uncertain payoff to obtain information) and exploitation (choosing policy that leads to higher payoff), and we can formulate this as a multi-armed bandit problem [56, 5, 13], treating each information policy as an arm. Formally, at each time $t = 1, ..., T$, ML chooses an information scheme τ_t, human with unknown but consistent behavior reacts to the information, and ML obtains a payoff $f(\tau_t)$ from human decisions. The goal of ML is to adaptively update the information policy to maximize the total payoff. The performance of bandit is measured in terms of *regret*, $\mathbb{E}[\mathsf{Reg}(T)] = Tf(\tau^*) - \sum_{t=1}^{T} f(\tau_t)$, where τ^* is the optimal solution of the problem (2) in hindsight assuming the human behavior model is known.

While the bandit formulation provides a nice foundation for our problem, the main challenge in our setting is that there are infinitely many information policies (i.e., infinitely many arms), and standard bandit algorithms could not converge to the optimal policy, i.e., it would not lead to *sublinear regret*. We plan to address this challenge using the technique in my prior work on adaptive contract design [47], in which we aim to find the optimal contract (among infinitely many of them) to crowd workers with unknown cost/effort levels by adaptively updating the contracts over time. The key intuition is that, when we select a contract, the response we obtained from humans provides us information about not only the posted contract but also other similar contracts (i.e., worker performance should be similar with similar payments). Therefore, we can propagate the information to nearby arms and achieve near-optimal learning. We plan to apply similar idea in information design through mapping a information policy to a contract. More formally, the feasibility of this learning problem hinges on the condition that posting similar information policy leads to similar payoff. Let $f(\tau)$ be the payoff after posting τ, as defined in optimization (2), let $D(\tau, \tau')$ denote the *distance* between two information policies. If we can find a Lipschitz constant L such that $|f(\tau) - f(\tau')| \leq L \cdot D(\tau, \tau')$ for

all (τ, τ'), we can adapt the techniques of my prior work to reach efficient learning. In this task, we will further tie this analysis to human models, i.e., characterizing situations for the Lipschitz condition to hold for different human models, deriving the corresponding Lipschitz constants, and developing bandit algorithms.

3.1.3 Task 1.3: Robust information design

We will also study settings in which we do not know the exact human model and we cannot learn from past interactions. The goal is to design an information policy that is *robust* to a range of possible human models.

Research questions. We plan to address this problem by borrowing ideas from robust contract design [15, 21, 69], in which robustness is defined using the notion of worst-case optimal, considering all the possible (even unknown) actions players can take. We will build a connection between a *contract* in contract design and an *information policy* in our setting. More specifically, since the main challenge is due to the uncertainty about humans, we can similarly define robustness based on the worst-case guarantee of human decisions (induced by human behavior models), over all possible decisions humans might take within a set of human models. Our goal is to design *robust optimal* information policy: the policy is *robust optimal* if the worst case performance is (weakly) better than that of all other policies. Consider the scenario where the human modeling comes from a function class $\mathcal{H}$, the robust-optimal policy with respect to $\mathcal{H}$ can be defined as

$$\max_{\tau \in \Delta(\Delta(\Theta))} \min_{(\omega(\cdot,\cdot),P(\cdot)) \in \mathcal{H}} \mathbb{E}_{\theta,\mu\sim\tau} \left[\sum_{a \in \mathcal{A}} P(a|\mu)) V(a,\theta) \right] \quad \text{s.t.} \quad \tau \in \mathcal{K} \tag{3}$$

My prior work [92] on robust learning, which utilizes the techniques from robust contract design to design robust decision rules for strategic users, will serve as the technical foundation for this task. The intuition of the technique is that, we can constrain the space of feasible policies from the partial knowledge we have, and search for the worst-case optimal within the constrained space. The key challenge in this setting is that we need to identify some sets of reasonable conditions that $\mathcal{H}$ should satisfy, e.g., increasing the payoff of a decision would increase the likelihood for that decision to be chosen, formulate the constraints those conditions introduce, and derive the robust optimal policy with respect to the different sets of conditions.

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

One key aspect in this research is to understand and characterize how humans make decisions in the context of machine-in-the-loop decision making. To formally understand the causal relationships of how different environment factors impact human decision making, we propose to conduct a series of large scale behavioral experiments. 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 study will examine whether and how humans deviate from being Bayesian rational in decision making, the common assumption in the literature, when machines are in the loop, and identify alternative models that better explain human behavior. We will also explore whether we can take interventions to make humans engage in more deliberations during decision-making, as this might sometimes be desired in decision-making with higher stakes or higher uncertainty. 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., having low generalization error for unseen data points. Therefore, we propose to utilize this connection and examine the expressiveness of human behavioral models through 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 online behavioral experiments, including examining whether crowd workers are rational towards financial incentives [45], how crowd workers are influenced by others [88, 28], and whether humans are Bayesian rational in the persuasion setting [87]. Here we describe the my most relevant work [87] that serves as the prelim-

inary result of the following proposed research. In particular, we focus on the persuasion setting [52] in which a sender aims to design information to persuade a receiver to take certain actions. We empirically examine that given prior beliefs, how receivers (i.e., humans) update their beliefs and take actions.

Experiment design: We recruited 400 workers from Amazon Mechanical Turk. Each worker is asked to complete questions involving probabilistic-inference and decision-making. 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, an urn will be secretly drawn according to the prior announced to workers, and then a ball will be randomly drawn from this urn. The color of the drawn ball will be disclosed to the worker. Upon seeing the ball color, a worker is asked to guess which urn is drawn and 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, differing 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 $(0.2, 0.8)$. We designed eight ball compositions in urns (information structures) such that, conditional on the realization of a red ball draw, the Bayesian posterior would be $(0.2, 0.3, \ldots, 0.9)$ for both treatments.

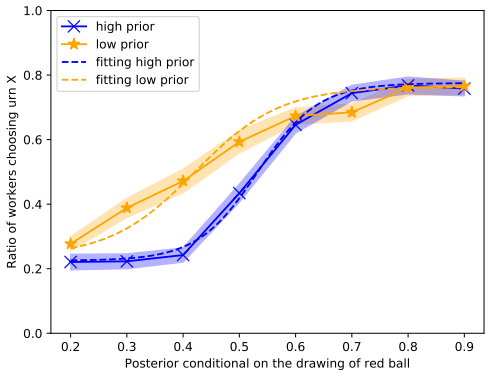


Figure 1: The solid lines represent the ratio of workers choosing Urn X. Shaded regions represent standard errors. Dashed lines represent fitted models using discrete choice model and probability weighting.

Experiment results: In our design, if workers are Bayesian, we should see no difference between two treatments, and if workers are rational, we should see workers choosing urn X when their posterior is greater than 0.5. Our results in Figure 1 show that workers have significantly deviated from being Bayesian rational. We also show that, an alternative human model (discrete choice model [67, 83, 94] coupled with probability weighting [100, 80, 75]) better aligns with workers' decision-making 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 examine human decision-making with machines in the loop under a wider range of scenarios. The goal is to understand the casual relationship between different factors (such as whether the information is provided by ML, the stake of the task, the impacts of prior, etc) and human decision-making process.

Experiments design. We will conduct a series of experiments under 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 human decision making. For the experiment design, the **independent variables** we plan to control across different treatments are (1) how we frame the information source, which could be from ML, humans, or a math expression (like the preliminary work above), to measure whether humans respond differently when there are machines in the loop, (2) different levels of *prior* and *information* (following the composition of Equation 1) to measure the impacts of belief updating, (3) different amount of payments to understand whether humans are more likely to be Bayesian rational when the stake is higher, and in addition, (4) different background scenarios and information presentation to understand how robust our observation is to different background contexts. For the **performance measure**, we will measure how well each decision model (belief updating plus decision-making) aligns with human behavior. We will start with the well-

known candidate models as surveyed in Section 2.2, however, we will introduce/include new models as the research evolves. Note that some human models are naturally more expressive (e.g., with more parameters). Directly measuring the fit to the data might be misleading. Since our goal is to understand which model *predicts* human behavior better, in this task, we will utilize the ML approach to measure how well each model aligns with human behavior, i.e., we will split the collected data into training and testing sets, use training set to fit the model parameters, and measure the accuracy of each model using test set.

Expected outcome. This task will provide us a comprehensive empirical understanding of real-world human decision-making process. In particular, we will obtain a series of causal relationships of how different factors impact human decision-making, measured by how well the models predict human behavior.

3.2.3 Task 2.2: Explore the impacts of interventions to human decision-making

The above task aims to understand and model real-world human behavior in machine-in-the-loop decision making. In this task, we consider the scenario in which we might want to take interventions to nudge humans to engage in more deliberations during decision-making, e.g., in scenarios when the stake of the decision or the uncertainty of the outcome is high. In particular, we plan to base our intervention design on the well-celebrated dual process theory (DPT) [33, 50, 19]. DPT specifies two processes through which human 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 during their decision making, it is believed that the default processing mode human brains would select is Type 1 processing. In this task, we aim to explore the design of interventions that nudge humans into more Type 2 processing in machine-in-the-loop decision making.

Design space of interventions. To design interventions to nudge humans into more Type 2 processing, there have been procedures developed in cognitive sciences [84, 99]. 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 of their decision 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. The interventions in this category could involve education, training, or guided meditation before decisions. (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 [14, 73], 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 [81, 62]. (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.

Experiment design 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 2.1 with the **independent variables** being the interventions. The **performance measure** will again be the prediction accuracy (in test set) of various behavioral models. This task will provide us an understanding of the effectiveness of different interventions in nudging humans towards more deliberations in their decisions, which could play an important role in scenarios when a certain decision has higher impacts to either the individual decision-maker or to the society.

3.2.4 Task 2.3: Quantify the expressiveness of behavioral models

In the above tasks, we aim to understand human decision-making process with machines in the loop under different scenarios and examine which behavioral models better explain human decision-making. To answer the questions, one common approach in the literature is to examine how well a model *fits* the data. However,

since a good behavioral model should be able to *predict* human behavior in the future, we adopt the ML approach with training/testing phases and use the accuracy on the test set as the measure. Looking at this problem through the lens of ML (i.e., considering behavioral models as ML hypotheses), in this task, we propose to investigate the problem of quantifying the expressiveness of human behavioral models.

Research questions. To illustrate the connection to ML, consider the following example on modeling human belief updating. Again, let $P(e)$ denote the prior of event e and I denote the signal/information. The Bayesian model predicts that humans update their beliefs following the Bayes rule $P(e|I) = \frac{P(I|e)P(e)}{P(I)}$. An alternative model could be to predict humans update their beliefs according to the rule $P(e|I) \propto P(I|e)^{\alpha}P(e)^{\beta}P(I)^{\gamma}$. To formulate this in ML terms, the prior $P(e)$ and signal distribution ($P(I|e)$ and $P(I)$) are the observable *features*, 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 examine whether humans are Bayesian, since there are 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 accuracy using standard concentration bounds, such as Hoeffding's inequality. However, when we are trying to fit the data to the alternative model, there are three variables (α,β, and γ) that we can tune to fit the data, which implies the expressiveness of the model 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.

The first natural attempt would be to characterize the expressiveness of behavioral models using the notion of VC dimension [96]. However, VC dimension is a measure that considers the worst-case data distribution. Since human behavior often follows some basic characteristics (e.g., more likely to choose a decision when the decision payoff is higher, the induced posterior is higher when the prior is higher), it is natural to consider a distribution-dependent approach. In this task, we will first identify some sets of reasonable axioms of human behavior and constrain our discussion in distributions that satisfy these axioms. We will then apply distribution-dependent notions to perform the analysis. In particular, we plan to start the investigation by adopting the *Rademacher complexity* [10, 55], which has two nice properties: (1) it is a distribution-dependent measure, and (2) it can be estimated with empirical data. We plan to analyze the Rademacher complexity with different set of axioms for human behavior. We will also empirically estimate the complexity using the data we collect in the experiments. This task will provide us a understanding of the capacities of human models in the ML perspective, which could shed lights in applying ML methods in behavioral theories, such as trading-off the approximability and generalizability of behavioral models and applying active learning to help design behavioral experiments to improve the efficiency of data collection.

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

We demonstrate that information design 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: how should we define the optimization objective? An ill-defined objective could pose potential concerns and lead to undesired outcomes. This problem is further amplified in high-stake domains with multiple objectives to balance. Take homelessness prevention for example, during decision making, homelessness service providers might want to reduce the number of families in need that are not offered assistance, reduce the expected number of families that would re-enter the system again in the future, and ensure the resource is allocated *fairly* to different social groups. Since these objectives do not always align, homelessness service providers often need to trade-off and balance these objectives. How should we take these trade-off into account and design appropriate objectives that align with human values? In this research thrust, we propose to address this concern by including human decision-makers in the loop to help determine the objective of the optimization through *eliciting and aggregating humans' decision criteria*.

This problem is domain-dependent in nature, and this thrust will build on my existing collaborations

with domain experts in applying computational approaches for societal problems, such as homelessness prevention [27] and kidney exchange [60]. In particular, we will work closely with Prof. Patrick Fowler (letter of collaboration attached) at Brown School of Social Work on applying the machine-in-the-loop decision making framework for homelessness prevention. We will also leverage the interdisciplinary efforts at WashU to expand the research to other domains. During the initial phases of the proposed research, we will shape and design the research questions with domain experts but will start by working with crowd workers from online crowdsourcing platforms. This approach enables us to perform more extensive examinations and gain a better understanding on the benefits and pitfalls of different approaches for eliciting and aggregating humans' decision objectives before deploying them in the field. With this understanding, we will then work with domain experts and field decision-makers to evaluate and deploy our approaches.

3.3.1 Task 3.1: Elicit individual decision criteria

In order to design objectives that align with human values, we need to first elicit the decision criteria of individual decision makers. There are two possible approaches for eliciting human decision criteria. The first one is to ask humans to report their criteria of decision making (more normative, eliciting what people should do), and the other is to infer the criteria from made decisions (more descriptive, eliciting what the criteria actually refers to in decisions). The are pros and cons in each approach. The first approach gives a direct answer to our question, but it is often hard for humans to accurately specify their decision criteria. The second approach lead to questions more natural for humans to answer but requires consistency in decision making. In this task, we will explore both approaches under different scenarios and investigate whether combining both approaches can support deliberation and lead to more consistent elicited reports.

Experiments. We will start the investigation by conducting the experiments on Amazon Mechanical Turk (and will work with decision makers in the field afterwards). After giving tutorials of the task background (e.g., homelessness prevention), workers are going to answer survey 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 infer the underlying criteria (again, in the form of ranking, hierarchical order, or weighted sum) that align with the made decisions. 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 a mixture, and different sets of expert-provided criteria. The reason to including answering a mixture of A and B as one of the treatments is for us to understand whether decision-makers form a more consistent decision-making policy through this process as it might encourage more deliberations. We will measure the results using distributions of answers (in survey A) and learned criteria (in survey B). We will also survey additional qualitative questions to gauge users' decision process and obtain feedback in the survey design.

Expected outcome. The results will help us understand both humans' criteria of decision making in the domain tasks and how different factors (e.g., familiarity of the task, elicitation methods, pre-provided set of criteria) impact the outcome of the elicitation. The results provide insights on human decision process and help us determine how to make the trade-off when deploying it in the field.

3.3.2 Task 3.2: Aggregate individual criteria into a collective objective

With the individual decision criteria elicited, our next task is to aggregate these criteria into an objective. 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.

Research questions. We will explore the aggregation with two different approaches and investigate their connections. In the first one, we will leverage social choice theory [4], 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). Given that it is often impossible to simultaneously satisfy all the desired properties during aggregation (e.g., Arrow's impossibility theorem [3]), different social choice mechanisms 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 a data point, defining the corresponding loss function, and finding an objective that minimizes the loss. In this task, we will examine the aggregation through the lens of the two approaches. In particular, since social-choice axioms can be viewed as constraints on the ML loss function (e.g., anonymity in social choice implies each point having the same weight in loss functions), we can explore questions such as, what social-choice axioms are satisfied for common loss functions in ML and how can we design loss functions to satisfy a set of given axioms, to develop a deeper understanding of the two approaches. In addition, we will recruit both domain experts and lay-persons (e.g., crowd workers) to obtain insights on which aggregation aligns more with what humans would do and also investigate the underlying rationales. The observations could in turn help us define axioms and loss functions during aggregation for the target domain application.

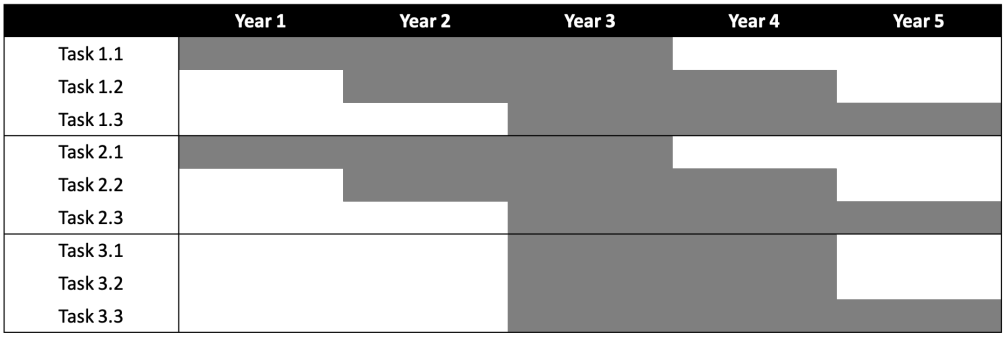
3.3.3 Task 3.3: Work with domain experts in real-world applications

We will work with domain experts to evaluate and deploy the results and findings in the previous tasks to real-world applications. The proposed activities will extend my existing collaboration with Prof. Patrick Fowler on designing algorithmic solutions for homelessness prevention [27] and aim to deploy the machine-in-the-loop approach for the problem. We will work with domain experts and local homelessness service providers, the St. Louis Area Regional Commission on Homelessness (SLARCH) – a nonprofit organization that coordinate homeless service provision across the St. Louis region, 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. We will also work with social workers, the decision makers in the field, recruited through SLARCH to evaluate and deploy our research.

In addition to homelessness prevention, we will 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 included below. The tasks Thrust 1 and 2 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 addresses the practical applications, after we have initial results for the first two thrusts.



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

- Data collection: The collected data of the behavioral experiments will be made publicly available to the research community. We believe the large-scale behavioral data would be of important research value.

- Modeling: We will examine user behavior models through prediction accuracy on the collected data. One of our research task (Task 2.3) also involves developing alternative measure for human modeling.
- 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.
- Deployment: We aim to deploy the proposed research in real-world applications. In addition to the evaluations above, we will work with domain experts, such as Prof. Patrick Fowler (support letter attached), to develop our evaluation plan and solicit feedback of the proposed framework through interviews/surveys.

4 Education Plan

The PI aims to broaden 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.

Outreach to high-school students. The PI will partner with the Institute for School Partnership (ISP) at WashU to design outreach activities for high-school students and teachers. The goal is to cultivate next-generation scientists/engineers through exposing high-school students to academic research and stimulating their interests in computing. We also plan to involve high-school teachers in the design and dissemination of the curriculum to maximize the outreach and impacts. Budgets are allocated for ISP for these activities (letter of collaboration 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 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 include a broad overview of machine learning (ML) and human behavior and engage students in group projects guided by Ph.D. students. We will prepare data sets and ML modules for students to 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 teacher 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 curriculum design to maximize the potential outreach and obtain feedback. We will then host the workshop in year 4 and 5 by recruiting around 20 high-school students with the help from ISP. Stipends for participating students are included in the budget: many low-income students rely on summer jobs to support their families and may not skip work for this camp. These stipends will allow the participating students to improve their learning ability compared to their high-income peers, without incurring a financial penalty on their families.

Evaluation plan: The ISP will help with the logistics of the workshop, including recruiting high-school teachers and students, and also provide consultations for evaluations. In particular, we will conduct anonymous surveys to high-school teachers/students before and after the event 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 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 [60] 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 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: *CSE 417T: Introduction to Machine Learning* and *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 available online to enable self-study or to be used in other institutions.

Synergy with research. The PI will deploy active learning techniques, such as peer instruction [20], in delivering the course. In fact, the research activities in this proposal have implications on how we shape teaching methods in education, in which the goal of the instructor (mapped to ML in the proposed research) aims to communicate with the students (mapped to human decision makers) to enable better learning. For example, my prior works [41, 88] 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 feedback, we could improve the long-term work performance of crowd workers. The results not only align with the observations of peer instruction [20] 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. 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. We have allocated budgets for the evaluation service. The evaluations will be conducted based on multiple metrics, including whether students obtain firm grasp of the subject (by constructing a knowledge inventory) and 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 gets 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 [27] and with Dr. Jason Wellen at Medical School to apply computational approaches for living donor kidney transplantation [60]. 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 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. We also develop a new foundational understanding of the legal landscape pertaining to fair decisions. The project so far has generated four publications [32, 71, 76, 92]. *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] Kenneth J Arrow. A difficulty in the concept of social welfare. *Journal of political economy*, 58(4): 328–346, 1950.

[4] Kenneth J Arrow. *Social choice and individual values*. Yale university press, 1951.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

[11] 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.

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

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

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

[19] 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.

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

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

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

[23] 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.

[24] 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.

[25] 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.

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

[27] 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.

[28] 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.

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

[30] 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.

[31] 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.

[32] Andrew Estornell, Sanmay Das, and Yevgeniy Vorobeychik. Incentivizing truthfulness through audits in strategic classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5347–5354, 2021.

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

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

[35] 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.

[36] 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.

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

[38] 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.

[39] 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.

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

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

[42] 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.

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

[44] 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.

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

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

[47] 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.

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

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

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

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

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

[53] 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.

[54] 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.

[55] Vladimir Koltchinskii and Dmitriy Panchenko. Rademacher processes and bounding the risk of function learning. In *High dimensional probability II*, pages 443–457. Springer, 2000.

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

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

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

[64] 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.

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

[66] 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.

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

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

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

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

[71] Quan Nguyen, Sanmay Das, and Roman Garnett. Scarce societal resource allocation and the price of (local) justice. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5628–5636, 2021.

[72] 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.

[73] 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.

[74] 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.

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

[76] Alexander Philipp Rader, Ionela G Mocanu, Vaishak Belle, and Brendan Juba. Learning implicitly with noisy data in linear arithmetic. In *4th Knowledge Representation and Reasoning Meets Machine Learning Workshop*, 2020.

[77] 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.

[78] 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.

[79] 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.

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

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

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

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

[84] 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.

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

[86] 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.

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

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

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

[90] 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.

[91] 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.

[92] 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.

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

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

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

[96] 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.

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

[98] 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.

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

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

[101] 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.

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

[103] 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.

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

[105] 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.