

Designing Behavior-Aware AI Systems to Influence
Human Decision-Making

Guanghai Yu

Ph.D. Dissertation. 2024

WASHINGTON UNIVERSITY IN ST. LOUIS

McKelvey School of Engineering
Department of Computer Science & Engineering

Dissertation Examination Committee:

William Yeoh, Chair
Chien-Ju Ho (Advisor)
Brendan Juba
Alvitta Ottley
Ming Yin

Designing Behavior-Aware AI Systems to Influence Human Decision-Making
by
Guanghai Yu

A dissertation presented to
the McKelvey School of Engineering
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2024
St. Louis, Missouri

© 2024, Guanghui Yu

Table of Contents

List of Figures	v
List of Tables	ix
Acknowledgments	xi
Abstract	xiv
Chapter 1: Introduction	1
Chapter 2: Background	6
2.1 Game Theory	6
2.2 Markov Decision Process	7
2.3 Techniques	9
2.3.1 Value Iteration	9
2.3.2 Policy Gradient	10
2.3.3 Proximal Policy Optimization	11
Chapter 3: Updating Decision-Making Environments	12
3.1 Related Work	14
3.2 Problem Formulations and Models	16
3.2.1 Environment Design Formulations	16
3.2.2 Reward Function Modification	19
3.2.3 Action Nudge	22
3.3 Experiments	24
3.3.1 Simulations	24
3.3.2 Human-Subject Experiments	26
3.4 Discussions	28
Chapter 4: Sending Persuasive Signals	30
4.1 Related Work	32
4.2 Problem Formulations and Models	34
4.2.1 Bayesian Persuasion Formulations	34
4.2.2 Encoding Human Behavior in Information Design	35
4.3 Experiments	39

4.3.1	Simulations	39
4.3.2	Human-Subject Experiments	44
4.4	Discussions	47
Chapter 5: Offering Predictions and Suggestions		50
5.1	Related Work	52
5.2	Problem Formulations and Research Questions	55
5.3	Experiments	58
5.3.1	Experiment 1: the Effect of Predictive Information	58
5.3.2	Experiment 2: the Effect of Prediction Sources	63
5.3.3	Experiment 3: the Effect of Value Similarity	66
5.4	Discussions	71
Chapter 6: Incorporating Human Beliefs about AI Behaviors		74
6.1	Related Work	77
6.2	Problem Formulations and Models	78
6.2.1	Decision-Making Environment	78
6.2.2	Modeling Human Behavior and Beliefs	79
6.2.3	Designing AI Agents	80
6.3	Experiments	81
6.3.1	Experiment Environments: Grid Worlds with Two Players	81
6.3.2	Experiment 1: Evaluating Human Behavior Models	83
6.3.3	Experiment 2: Evaluating Human Belief Models	84
6.3.4	Experiment 3: Designing Collaborative AI Agents	87
6.4	Discussions	89
Chapter 7: Conclusion and Future Directions		90
References		94
Appendix A: Proofs of Theorems		111
A.1	Proof of Lemma 1	112
A.2	Proof of Theorem 2	113
A.3	Proof of Lemma 3	115
A.4	Proof of Lemma 4	119
Appendix B: Supplemental Experiment Results		121
B.1	Additional Experiment Results in Chapter 3	122
B.2	Additional Experiment Results in Chapter 4	125
B.2.1	Data Generation in Experiments	125
B.2.2	Convergence and Scalability of Proposed Methods	126
B.2.3	Generalizability of Proposed Methods	128
B.3	Additional Experiment Results in Chapter 5	132
B.3.1	the Effect of Prediction Magnitude	132

B.3.2	the Effect of Value Similarity Claims	132
B.4	Additional Experiment Results in Chapter 6	135
B.4.1	Experiment Environments: Grid Worlds with Single Player	135
B.4.2	Experiment 4: Evaluating Human Behavior in Single-Player MDP	135
B.4.3	Experiment 5: Evaluating Human Belief Models	136
B.4.4	Experiment 6: Evaluating Collaborative AI Agents	138
Appendix C:	Details of Human-Subject Experiments	140
C.1	Human Experiment Details in Chapter 3	141
C.2	Human Experiment Details in Chapter 4	142
C.3	Human Experiment Details in Chapter 5	145
C.4	Human Experiment Details in Chapter 6	145
C.5	Demographic Information of Human-Subject Experiments	148

List of Figures

Figure 1.1:	The decision-making framework discussed in this dissertation involves the decision maker gathering information from the environment and taking actions that subsequently update the environment. We explore methods to modify the environment (in Chapter 3), select relevant information (in Chapter 4 and Chapter 5), and design AI teammates (in Chapter 6) for the decision-maker, with the goal of influencing the decisions made by real humans.	3
Figure 3.1:	The principal’s payoff with biased decision-makers without environment design. Agents with higher τ or lower k are closer to being rational.	25
Figure 3.2:	The principal’s payoff with biased decision-makers after applying environment design. The y-axis is the relative performance compared with the optimal solutions, and the x-axis is the amount of budget spent relative to the optimal performance.	25
Figure 3.3:	Misalignment of the principal’s and the agent’s the agent’s reward function. The y-axis is the relative performance compared with the optimal performance (in terms of the principal’s payoff), and the x-axis is the amount of budget spent relative to the optimal performance.	26
Figure 3.4:	The human-subject experiment results of environment design. The results are grouped by the vision length of the games, mapping to different values of τ in short-sighted (boundedly rational) agents. Figure 3.4a shows the average principal’s payoff with real human decision makers in treatments, and Figure 3.4b shows the ratio of worker moves which are the same as short-sighted model predictions.	28
Figure 4.1:	The HAIDNet framework. The human descriptor module is given to the optimization module before training. The optimization is performed through back propagation which evaluates the gradient of the loss to update the weights in the neural network structure.	36

Figure 4.2:	Comparing the optimal information policy and the policy generated by HAIDNet in the setting with binary actions and binary states.	41
Figure 4.3:	The performance of HAIDNet in settings when the receiver is not Bayesian rational. We train HAIDNet with non-Bayesian-rational receiver model parameterized by β_H , then evaluate the learned information policy for all receiver models. The performance is normalized so for each human model, the optimal performance is 1.0 among all policies.	44
Figure 4.4:	Average sender utility of different policies in human-subject experiments of Phase 2. The differences between BR-policy and TH-policy and between BR-policy and HAIDNet are statistically significant ($p < 0.01$).	47
Figure 5.1:	Human experiment interface of ethical decision-making in kidney allocation with AI generated predictive information. The information of post-transplant survival chance is generated by AI systems.	57
Figure 5.2:	The effect of equal prediction on human decisions. We present ΔP for each verifiable factor and treatment. There is no significant difference between treatments in the Prior Donor factor ($p = .54$). There is a significant difference between treatments in the Wait Time factor ($p = .045$). There is a significant difference between treatments in the Disease Stage factor ($p = .0057$).	61
Figure 5.3:	The effect of aligned prediction on human decisions. There is a significant difference between a misaligned prediction and equal prediction for all factors ($p < .001$). There is a significant difference between a equal prediction and aligned prediction for all factors ($p < .003$).	63
Figure 5.4:	The effect of prediction source on human decisions. There is no significant difference between treatments in the Prior Donor factor, Wait Time factor, or Disease Stage factor. There is a significant difference between treatments in the Predictive factor ($p = .0316$).	64
Figure 5.5:	The effect of value similarity on alignment change between Stage 1 and 2. In the left figure, we find across all scenarios, the dissimilar AI has a significantly larger change in alignment ($p < .001$). In the right figure, we find that in scenarios where the human and AI disagree, the similar AI has a significantly larger change in alignment ($p = 0.003$).	70

Figure 6.1:	An example task of encoding beliefs about AI models into human-AI cooperation problem.	75
Figure 6.2:	The interface of human-subject experiments in human-AI cooperation. In Experiment 1, each participant is playing with an AI agent. The participants in experiment 1 are told what their goal is and only need to focus on reaching the goal without colliding with the AI agent. In Experiment 2, each participant is provided traces of the behavior by other agents and is asked to infer which goal one agent is trying to reach. Finally, in Experiment 3, the participants are not told which goal to reach, and they need to make decisions based on their beliefs over AI behavior. In Experiments 1 and 3, the participants can only receive bonus rewards by reaching the same type of goals (star or triangle) as the AI agent. . .	83
Figure 6.3:	Belief inference results in simulations and human-subject experiments in Experiment 2.	86
Figure 6.4:	The average human-AI collaboration performance in human-subject experiment in Experiment 3.	88
Figure A.1:	The example MDP used for proving Lemma 1 with bounded-rational agents.	112
Figure B.1:	Examining the impact of β to the relaxed reward function modification algorithm. When β is large, the relaxation is close to optimal. The results suggest that a small β is sufficient for the approximation.	123
Figure B.2:	Performance of reward modification and action nudge methods in environment design problems when the rewards and bias parameters are inferred from observations.	124
Figure B.3:	The convergence results, with respect to the number of training iterations and β , of the sender’s utility derived from the information policy generated by HAIDNet.	127
Figure B.4:	Effect of prediction magnitude in Experiment 1. We present ΔP for each magnitude of prediction difference in blue, and ΔP for the verifiable only treatment group in black.	133

Figure B.5: The effect of value similarity on alignment change between Stages 1 and 2 in Experiment 3, across combinations of Deterministic/Random and Similar/Dissimilar treatments. When the AI is Deterministic, the Similar AI leads to a significantly larger change in conditional alignment ($p < .001$). However, when the AI is Random, there is no significant difference between Similar and Dissimilar AI ($p = .58$).	134
Figure B.6: Human-subject experiment interfaces. In Experiment 4, each participant is asked to control the player to move to the goal (red star). In Experiment 5, each participant is provided a trace of the behavior by another agent, and is asked to infer which goal the agent is trying to reach. In Experiment 6, each participant is playing with an AI agent in separate environments.	136
Figure B.7: Human evaluations of belief inference accuracy regarding AI goals. . . .	137
Figure B.8: Average collaborative reward of humans and AI agents in Experiment 6.	139
Figure C.1: Human experiment interface of updating decision-making environments in Chapter 3. Workers can use arrow keys to move the airplane around and collect points. The information on the bottom of the left-panel is the action nudge presented to workers, which is shown in action nudge treatment when a nudge is provided by AI model, hidden otherwise. . .	142
Figure C.2: Human experiment interface of designing information policy in Chapter 4.	144
Figure C.3: Human experiment interface of human ethical decision-making in Chapter 5.	146
Figure C.4: Human experiment interfaces of goal navigation and belief inference tasks in Chapter 6.	147

List of Tables

Table 4.1:	Comparing the average sender utility generated by the optimal policy and the policy from HAIDNet in the setting with a single Bayesian rational receiver.	42
Table 4.2:	Comparing the average sender utility generated by the optimal policy and the policy from HAIDNet in the setting with K Bayesian rational receivers.	43
Table 4.3:	Comparing the average sender utility by the optimal policy and the policy from HAIDNet in the setting with a non-Bayesian-rational receiver parameterized by β_H	45
Table 4.4:	Test accuracy of different human behavior descriptors in human-subject experiments of Phase 1.	46
Table 6.1:	The prediction accuracy for human behavior assuming optimal behavior and using data-driven model in Experiment 1.	84
Table 6.2:	Simulation results of collaborative performance over 10k testing cases in Experiment 3. Column players are different AI agents, and row players are different simulated human models.	88
Table B.1:	Comparing run-time between HAIDNet and linear programming methods. K is the number of receivers. The reported run-times are in seconds.	128
Table B.2:	Comparing the average sender utility generated by the optimal policy and the policy from HAIDNet in the setting with a single Bayesian rational receiver.	129
Table B.3:	Comparing the average sender utility by the optimal policy and the policy from HAIDNet in the setting with at most 10 Bayesian rational receivers.	130

Table B.4:	Comparing the average sender utility by the optimal policy and the policy from HAIDNet in the setting with at most 5 states and 5 actions, for a single Bayesian rational receiver.	130
Table B.5:	The prediction accuracy for human behavior for different human models in Experiment 4.	136
Table B.6:	Performance comparison between Bayesian inference framework using standard model and human behavior model.	137
Table B.7:	Simulation results of human-AI collaborative rewards in Experiment 6. Columns players are different AI agents, and row players are different simulated human models.	139
Table C.1:	Comparing sender and receiver utility of different policies in human-subject experiments of designing information policy.	143
Table C.2:	Task setup of each human-subject experiment.	148
Table C.3:	Demographic information of all the participants in our human-subject experiments.	149

Acknowledgments

I would like to express my deepest gratitude to my advisor, Chien-Ju Ho (CJ). CJ has provided immersive guidance not only in my academic research but also in my personal development and broader life, greatly enriching my study experience. CJ has granted me the freedom to explore various research topics and encouraged me to gain industry experiences. He is the ideal advisor for a young scholar pursuing a PhD degree.

I am profoundly thankful to my committee members: William Yeoh, Ming Yin, Brendan Juba, and Alvitta Ottley. I want to thank William for his insightful feedback and valuable suggestions on my dissertation, as well as his engaging discussions on our goal recognition work. I'm grateful to Ming for her deep understanding of ethical AI and human decision-making, which significantly influence my research, and also for her guidance on our work about AI's influence on human decision-makers in ethical decision domains. My thanks also go to Brendan and Alvitta for their continued support and feedback on my proposal and dissertation.

I want to show my thanks to my collaborators and co-workers, who are pivotal to my research and study. I thank Wei Tang for his insights during my early PhD years and ongoing support throughout the whole journey. His shared knowledge and experiences greatly illuminated my PhD study and helped me avoid many detours. I am indebted to Saumik Narayanan for sharing his wisdom across various domains and supporting the execution of experiments. Discussions with him always spark innovative ideas for our research. I want to thank Robert Kasumba for his insightful discussions and extensive collaboration in human-AI collaboration work. I also appreciate Alex DiChristofano, Lauren Treiman, and Tory Farmer for their shared insights on broader topics, rehearsal sessions for numerous presentations, and their valuable feedback. They have shown me how to be a good PhD student in every aspect and have brought many unforgettable memories outside the campus.

Finally, I must thank my family and friends for their support throughout the challenging PhD journey. The journey has brought many challenges beyond research, often resulting in unexpected and uncontrollable situations. I am so lucky to always have people there to guide me back to the right path. I am grateful to everyone who has helped, supported, and

crossed paths with me during this unique experience. Thank you for all your companionship, help, encouragement, advice, and surprises throughout my life.

Our work was supported in part by the Office of Naval Research Grant N00014-20-1-2240, the NSF FAI program in collaboration with Amazon under grant IIS-1939677, a J.P. Morgan Faculty Research Award, and a Global Incubator Seed Grant from McDonnell International Scholars Academy.

Guanghai Yu

Washington University in St. Louis
August 2024

To my grandfather.

ABSTRACT OF THE DISSERTATION

Designing Behavior-Aware AI Systems to Influence Human Decision-Making

by

Guanghai Yu

Doctor of Philosophy in Computer Engineering

Washington University in St. Louis, 2024

Professor Chien-Ju Ho, Advisor

Artificial intelligence (AI) demonstrates superiority over humans in many applications, such as image processing, speech recognition, and decision-making. AI systems are more efficient and accurate in information processing and computation, making them invaluable in assisting human decision-making across domains like healthcare, finance, and academia. Despite these strengths, real-world applications often involve complex, context-specific judgments and a deep understanding of human values, which current AI systems might not fully grasp. In domains requiring creativity and critical thinking, AI systems heavily rely on existing human knowledge. Humans, however, possess unique strengths such as intuition, which aids decision-making in uncertain or novel situations, and moral reasoning, which transcends mere efficiency or optimization. Therefore, combining the strengths of both humans and AI systems is essential for effective decision-making in real-world applications.

While AI systems offer significant advantages in supporting human decision-making, their incorporation poses considerable challenges. Human decision-making is complex and characterized by numerous unique traits. Empirical studies reveal that humans can be irrational, unconscious, and sometimes unpredictable, exhibiting various biases such as framing effects and confirmation bias. To develop AI systems that can interact effectively with humans, it is crucial to consider these human traits. However, the integration of human decision models

into AI systems has not been sufficiently explored in the field of human-AI interaction. Understanding and predicting human behavior and beliefs can enhance AI’s ability to support and influence human decision-making, leading to more effective and human-aligned outcomes. By incorporating human decision models, AI systems can provide more personalized and context-aware assistance, improving both decision quality and human satisfaction.

In this dissertation, we address the characteristics of human decision-makers and explicitly incorporate human models into algorithm design. Our study aims to model human behavior and beliefs about AI systems in both one-shot and sequential decision-making scenarios. By leveraging existing research in psychology and economics, as well as data-driven methods using data collected from real humans, we develop models that predict actions more accurately than those based on the assumption of human rationality. We also construct belief models to describe how humans adjust their actions in response to other players within the same decision-making environment. Utilizing these human models, we explore strategies to assist or influence human decisions by designing information signals, modifying decision-making environments, and developing AI teammates that operate alongside humans. Real human-subject experiments are conducted to gain a deeper understanding of human behavior in specific applications and to validate that our designed AI systems can effectively interact with human decision-makers, guiding their decisions toward predefined goals.

Chapter 1

Introduction

Artificial Intelligence (AI) has made significant changes in our daily life. In game playing, AI can achieve better performance than even the most talented human player; for example, DeepMind’s AlphaGo beat the world champion in 2016. AI models can help humans conduct dangerous missions or perform simple repetitive tasks, such as Boston Dynamics’ robot dogs utilized in search and rescue missions. Reinforcement learning algorithms can solve complex tasks when a proper reward function is defined, and recommender systems might know users’ preferences better than they do themselves, thanks to large-scale data mining techniques. Large language models (LLMs) can pass some professional and academic exams (e.g., GPT-4 scored in the top 10% in GRE Verbal and AP Statistics [2]), and even possibly outperform human experts in some domains [188, 114]. Compared with human decision-makers, AI algorithms show advantages in accuracy, scalability, and speed, so they are deployed to assist human decision-makers or even significantly influence their decisions.

However, AI systems are not yet ready to replace human decision-makers in every domain. AI systems can exhibit biases in their decisions, such as in healthcare (for example, [133] found that computer-aided diagnosis (CAD) systems have lower accuracy for minority groups than for the majority), hiring, and image generation. They might violate privacy requirements or intellectual property rights and raise ethical concerns in high-stakes domains when their decisions are not aligned with human values. It is often challenging to design AI systems that align with human-intended goals. Many advanced models rely on a predefined loss function, which can be challenging or even impossible to define in some cases (for example, designing an AI system to solve the problem of Goldbach’s conjecture). Therefore, it is necessary to involve both humans and AI systems in the decision-making process to leverage the advantages of both.

Incorporating human models into algorithm design could make AI more accurate, efficient, and transparent when humans are in the loop. Human decision-making models include personal utility, preferences, beliefs, intents, and other external factors such as social norms and cultural differences. It is hard to precisely capture human behavior because real humans are irrational, unconscious, and often unpredictable. Previous work has investigated the characteristics of the human decision-making process and proposed models to describe specific types of biased behaviors. For example, the hyperbolic discounting factor [134] is proposed to capture the inconsistency and disproportion of human discounting of short-term rewards and long-term rewards. Even without prior knowledge, it is still feasible to collect human behavior and responses, and then build human models using a data-driven approach [89]. Whether from prior knowledge or through a data-driven approach, realistic human models are far from the rational assumption (maximizing expected utility), which changes the optimization problem for designing AI systems to interact with human decision-makers.

This dissertation examines the interactions between AI systems and human decision-makers, focusing on scenarios where human decisions will result in rewards for another party. For example, when customers purchase items online, the shopping website earns profits. The website can design its layout, modify product information, and adjust item prices to attract potential customers. In this context, the shopping website is referred to as the principal, defined as the party or individual with the ability and power to influence human decision-makers. The principal aims to utilize AI systems (e.g., the recommendation system used by the shopping website) to guide human decisions in alignment with its preferences.

Figure 1.1 provides a general decision-making framework used in this dissertation. The decision-maker collects information from the decision-making environment, and their decisions subsequently update the environment. AI systems can be deployed into almost every aspect of this decision-making framework to influence human decisions and their outcomes. To design AI systems that can effectively interact with human decision-makers, we begin by modeling human behavior in both one-shot and sequential decision-making environments. We propose both method-based and data-driven approaches to build human behavior models, demonstrating their superiority in predicting human actions over commonly assumed rational models. Using these human models, we explore strategies to influence decisions by designing information signals, modifying decision-making environments, and developing AI teammates that act concurrently with humans.

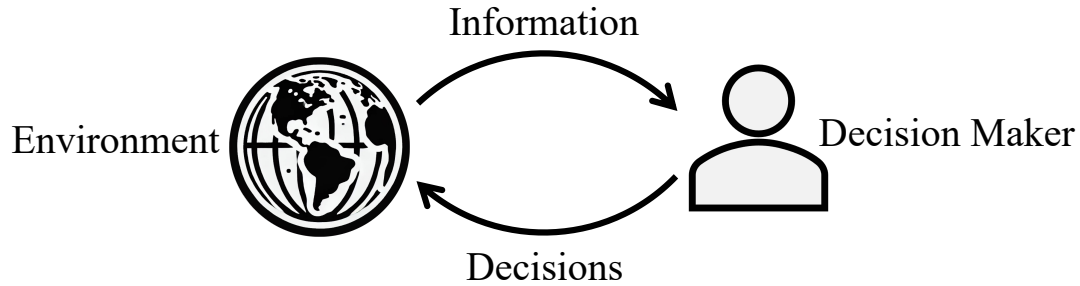


Figure 1.1: The decision-making framework discussed in this dissertation involves the decision maker gathering information from the environment and taking actions that subsequently update the environment. We explore methods to modify the environment (in Chapter 3), select relevant information (in Chapter 4 and Chapter 5), and design AI teammates (in Chapter 6) for the decision-maker, with the goal of influencing the decisions made by real humans.

The dissertation is organized as follows:

In Chapter 2, we discuss the decision-making problem setup in both one-shot games and sequential cases. We introduce the common assumption of rationality or optimal behavior, as well as some state-of-the-art methods to solve those decision-making problems, which will be used as baselines in our study.

In Chapter 3, we discuss how AI systems could update the decision-making environment to induce humans to take certain actions or make (near) optimal decisions in sequential decision-making problems. We relax the common assumption that the decision-maker is rational and incorporate a biased human model into the environment design problem. We propose two design methods for AI models and run experiments to evaluate human decisions influenced by the proposed AI systems. Our findings highlight the need to incorporate realistic human behavior models when designing AI systems to work with real humans. This Chapter is based on

Yu, G., & Ho, C. J. (2022). Environment Design for Biased Decision Makers. In IJCAI (pp. 592-598).

In Chapter 4, we investigate what information an AI system should present to the decision-maker. We seek the answer by framing the question as an information design problem. In

information design, a *sender* (the AI system) aims to design an information disclosure policy to influence the *receiver* (the human decision-maker) in decision-making. We propose a general architecture to solve this problem for various behavior patterns beyond the standard Bayesian rational assumption. To accurately model human decision-makers, we utilize a similar method-based approach as in Chapter 3, and we also explore some data-driven method and find the data-driven method could predict human actions more accurately than the method-based approach in our experiment setup. This Chapter is based on

Yu, G., Tang, W., Narayanan, S., & Ho, C. J. (2023). Encoding human behavior in information design through deep learning. In Proceedings of the 37th International Conference on Neural Information Processing Systems (pp. 7506-7528).

In Chapter 5, we expand our research beyond the optimization of pre-defined utility functions in AI systems to address ethical decision-making challenges. Here, AI systems are tasked with assisting human decision-makers in the allocation of medical resources. We examine the impact of predictive information, the sources of these predictions, and the alignment of values between AI systems and humans on human ethical decision-making. Our findings indicate that predictive information significantly influences human ethical preferences during the decision-making process. This work is based on

Narayanan, S., Yu, G., Tang, W., Ho, C. J., & Yin, M. (2022). How Does Predictive Information Affect Human Ethical Preferences?. In Proceedings of the 2022 AAI/ACM Conference on AI, Ethics, and Society (pp. 508-517).

and

Narayanan, S., Yu, G., Ho, C. J., & Yin, M. (2023). How does Value Similarity affect Human Reliance in AI-Assisted Ethical Decision Making?. In Proceedings of the 2023 AAI/ACM Conference on AI, Ethics, and Society (pp. 49-57).

In Chapter 6, we explore the design of AI teammates as decision-makers that influence human decisions. Rather than having humans as the sole decision-makers, an AI teammate collaborates with them in real times, with joint rewards determined by the actions of both humans and AI systems. This cooperation introduces new challenges related to human dynamic behavior. To address these challenges, we consider the beliefs of human players and

how they perceive the behavior of their AI partners. We train AI teammates to act in ways that reveal their action plans, thereby facilitating smoother collaboration with their human counterparts. We design various two-player cooperation games to assess the effectiveness of human-AI collaboration. This Chapter is based on

Yu, G., Kasumba, R., Ho, C. J., & Yeoh, W. (2024). On the Utility of Accounting for Human Beliefs about AI Behavior in Human-AI Collaboration. Under review.

We summarize the dissertation in Chapter 7, and discuss limitations of our work and some future directions along this line of research. We include proofs of theorems, extensive experimental results, and details of our human-subject experiments in the Appendix.

Chapter 2

Background

In this chapter, we present the background and an overview of the techniques employed in this dissertation. We begin by discussing the general problem settings for one-shot and sequential decision-making problems, as well as models of rational (or optimal) decision-makers. Following this, we delve into the techniques used to design AI systems, which will serve as baselines in the following chapters.

2.1 Game Theory

Game theory is a mathematical framework that investigates the strategic interactions of multiple agents and the outcome of their decisions [207, 186]. It's prevalent in numerous fields, such as economics, business, political science and computer science. In a one-shot game with single player, we define the action space as $a \in A$, and utility function $U : A \rightarrow \mathbb{R}$. For games with n players, we define the joint action space as $A = A_1 \times \dots \times A_n$ and reward function $U_i : A \rightarrow \mathbb{R}$ for each player $i \in [n]$. Here A_i is the action space of the i -th player, and U_i is the utility function of the i -th player. The notation (a_i, a_{-i}) refers to the action taken by the i -th player and the actions of all other players except the i -th player.

Rational agents. The assumption of rational agents is widely adopted in game theory. All players in a game are usually assumed to be rational, meaning they consistently act in a way that maximizes their own utility based on their preferences and knowledge of the game. This assumption implies that agents can reason about the possible outcomes of their choices and select the one that best serves their interests or matches their preferences. Additionally, rational agents can reason about the actions of other players. Key concepts such as Nash Equilibrium are derived from the rationality assumption. In Nash Equilibrium, each player's

strategy is optimal given the strategies of others, and no player has an incentive to deviate from their chosen actions.

We model the rational/optimal agent as a player who always chooses actions to maximize their utility. That is, in one-shot game, a rational agent will always choose action to maximize their utility, $a^* = \operatorname{argmax}_a U(a)$. For multi-player games, we take the notion of Nash Equilibrium. Optimal agents will take a_i^* which satisfies $U_i(a_i^*, a_{-i}^*) \geq U_i(a_i, a_{-i}^*), \forall i \in N$, meaning no agent could gain more utility via changing their actions. In the scope of this dissertation, we simply assume the existence of at least one equilibrium, although this is not always true in the general case.

Level-k reasoning. In multi-player games, level- k reasoning describes how players make decisions based on their predictions of the likely actions of other players. A level-0 agent ignores the actions of other players and takes actions to maximize their own utility. A level- k player assumes all other players are behaving as level- $(k - 1)$ players, and thus their optimal policy is the best response to level- $(k - 1)$ players. For player i at level- k , their action is denoted as $a_{i,k}^* = \operatorname{argmax}_{a \in A_i} U_i(a, a_{-i,k-1}^*)$, where all other players are assumed to take their actions as level- $(k - 1)$ agents, $a_{-i,k-1}^*$. The optimal solution (which can be derived using the definition of Nash Equilibrium) can be viewed as all players taking level- ∞ actions.

Biased agents. The rationality assumption may not hold when decision-makers are real humans. Previous research in psychology and economics has shown that humans often deviate from being rational in multiple cases. To model biased decision-makers, we utilize both model-based methods (such as hyperbolic discounting factors and discrete choice models) and data-driven approaches using real human datasets collected from designed experiments. We will discuss the details of biased decision-maker models in the following chapters.

2.2 Markov Decision Process

For sequential decision problems, a Markov decision process (MDP) is a common approach to describe the problem setup. Here, we introduce the formulation of MDPs and techniques to find optimal solutions. We will extend the discussion to cases involving irrational or biased decision-makers and explore methods to address these challenges later.

MDP formulations. A standard formulation of MDP is defined as $W = \langle S, A, P, R \rangle$, where S is the set of states, A is the set of agent actions, $P(s'|s, a)$ is the transition probability from state s to state s' after taking action a , and $R(s, a)$ is the bounded reward obtained by the agent after he takes action a at state s . The transition dynamics describe the key feature of MDP, the Markov property, which states that the future state will only depend on the current state and the action taken, while being independent of the action or state history.

Optimal policy. The agent aims to find a (stochastic) policy $\Pi : S \times A \rightarrow [0, 1]$, specifying the probability of actions to be taken in each state. The objective is to maximize the cumulative reward over either a finite or infinite time horizon. The cumulative reward is defined as shown in Equation 2.1, where T is the maximum decision time in cases of finite time horizon, and $0 < \gamma \leq 1$ is the discount factor for cases of infinite time horizon. In this formulation, (s_t, a_t) represents the state and action at time t , the initial state is given as s_0 , and $s_{t+1} \sim P(s|s_t, a_t)$ is drawn from the transition dynamics.

$$\mathbb{E} \left[\sum_{t=0}^T R(s_t, a_t) \right] \quad \text{or} \quad \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \right] \quad (2.1)$$

Multi-agent MDP. When multiple decision-makers are available, we could extend the formulation into a multi-agent case, represented as $W = \langle S, \alpha, A_{i|i \in [\alpha]}, P_{i|i \in [\alpha]}, R \rangle$, where α is a finite set of players; A_i is the action set available to player i . $P : S \times A_1 \times \dots \times A_{|\alpha|} \times S \rightarrow [0, 1]$ is the transition function that determines the next state given all players' joint actions. $R_i : S \times A_1 \times \dots \times A_{|\alpha|} \rightarrow \mathbb{R}$ is the reward function assigned to player i given their joint actions.

The challenge brought by multi-agent MDP is that when each agent optimizes their own policy without considering the actions of others, it can disrupt the Markov property, making solutions non-stationary or unstable. Some techniques have been proposed to solve this challenge in some special cases, such as the two-player adversarial case and the multi-player cooperative case. In this work, we discuss a special two-player cooperation case in Chapter 6.

2.3 Techniques

In this dissertation, we use value iteration, policy gradient, and their variants to solve the sequential decision-making problems. Here, we introduce the standard techniques to solve MDPs with an infinite time horizon, and these approaches could be applied to the finite time horizon case as well.

2.3.1 Value Iteration

Before we introduce value iteration, we introduce the concept of the Bellman Equation, which decomposes solving the optimization problem into simpler sub-problems. It introduces a value function $V(s)$, which is the expected cumulative reward starting from the initial state. The optimal value function V^* satisfies Equation (2.2).

$$V^*(s) = \max_a \left[R(s, a) + \gamma \sum_{s'} P(s'|s, a) V^*(s') \right] \quad (2.2)$$

We derive the value iteration method from the Bellman Equation. Value iteration is known for its simplicity and effectiveness in finding the optimal policy for finite MDPs. However, it can be computationally intensive for large state spaces, as it requires updating the value for each state at every iteration. Despite this, value iteration remains a fundamental algorithm in reinforcement learning, providing a clear and direct approach to solving MDPs. As shown in Equation (2.3), we iteratively update the value function and Q -values until convergence to find the optimal solution a^* at each state. The convergence properties and theoretical guarantees of value iteration make it a reliable method for policy determination.

$$\begin{aligned} a^* &= \operatorname{argmax}_a Q(s, a) \\ V(s) &= \max_a Q(s, a) \\ Q(s, a) &= r(s, a) + \gamma \sum_{s'} P(s'|s, a) V(s') \end{aligned} \quad (2.3)$$

For problems with large or continuous action spaces or state spaces, where the table representation of Q values might not be feasible, methods such as Deep-Q Network (DQN) have been proposed to solve the problem.

2.3.2 Policy Gradient

Policy gradient methods are a class of reinforcement learning algorithms that optimize the policy directly, rather than estimating value functions. These methods are particularly useful in environments with continuous action spaces or where the policy needs to be represented by complex functions, such as neural networks. The key idea is to parameterize the policy $\pi_\theta(a|s)$ with a set of parameters θ (for example, the weights of a neural network), and optimize the policy parameters using the gradient of the expected return J as shown in Equation (2.4). The policy gradient theorem provides the foundation for these methods (the proof can be found in [170] Section 13.2), where $Q^\pi(s, a)$ is the Q-value under policy π .

$$\begin{aligned}
 J &= \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, \pi_\theta(a|s)) \right] \\
 \nabla_\theta J(\theta) &= \mathbb{E}_{\pi_\theta} [\nabla_\theta \log \pi_\theta(a|s) Q^\pi(s, a)]
 \end{aligned}
 \tag{2.4}$$

Policy gradient methods, such as REINFORCE and Actor-Critic algorithms, leverage the gradient $\nabla_\theta J(\theta)$ to iteratively update the policy parameters in a direction that increases the expected return. These methods are powerful for their flexibility and capability to handle high-dimensional and continuous action spaces.

Actor-Critic. Actor-Critic methods combine both policy-based methods (the Actor) and value-based methods (the Critic). The actor will learn a policy to make a decision, and the critic evaluates the actions taken by the actor. In our work, we use neural networks of the same structure to represent the actor and the critic respectively, and optimize the neural network parameters with gradient methods.

2.3.3 Proximal Policy Optimization

Traditional policy gradient methods, while effective, often suffer from high variance and instability during training. [157] introduce a variant of policy gradient methods called Proximal Policy Optimization (PPO), which utilizes a surrogate objective function that ensures the new policy does not deviate too much from the old policy. This is achieved by constraining the policy update to stay within a small, trust region around the current policy, thereby preventing large, potentially harmful updates. The policy objective function is shown in Equation (2.5):

$$\begin{aligned}\hat{A}_t(s, a) &= R(s, a) + \gamma \sum_{s'} V(s') P(s'|s, a) - V(s), \\ L^{CLIP}(\theta) &= \mathbb{E}_t \left[\min \left(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right],\end{aligned}\tag{2.5}$$

where:

- $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ is the probability ratio between the new policy and the old policy.
- \hat{A}_t is the estimated advantage function at time step t .
- ϵ is a hyperparameter that controls the clip range.
- $\text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)$ constrains the probability ratio to lie within the range $[1 - \epsilon, 1 + \epsilon]$.

The clipping mechanism ensures that the update is conservative by limiting the extent to which the policy can change at each step. This helps maintain a balance between exploration and exploitation, reducing the risk of catastrophic performance drops. Due to its robustness, ease of tuning, and strong empirical performance, we use this method as baselines to develop our AI systems.

Chapter 3

Updating Decision-Making Environments

In this chapter, we examine how AI systems can modify the decision-making environment to encourage humans to take specific actions or make (near) optimal decisions in sequential decision-making problems. Given that real humans may exhibit biases during the decision-making process, it is essential to account for these biases when designing AI systems. Our findings indicate that ignoring these biases can render AI systems ineffective, whereas incorporating prior knowledge of these biases can maintain the effectiveness of the AI systems. This chapter is based on joint work with Chien-Ju Ho [203]. I contributed to this work by framing the problem formulation, proposing solutions, and conducting both simulations and human-subject experiments.

To study this problem, we consider instances where AI systems and human decision makers might have mis-aligned objectives. More specifically, AI systems are playing as a *principal*, and human decision makers are playing as an *agent*, in the same sequential decision making environment. The goal of the agent is to take a sequence of actions to maximize his total payoff¹. The principal cannot directly take actions but can update the environment to influence the agent's actions and receive reward based on the agent's actions. The goal of the principal is to update the environment such that the agent takes actions that maximize the principal's payoff.

This problem setting is motivated by several existing and potential applications. For example, a user-generated content website might want to update their site to provide incentives, such as badges or virtual points, to encourage users to consume and rate the content on their website. An online retailer might want to decide when and whether to provide coupons to

¹We use *she* to denote the principal and *he* to denote the agent respectively.

nudge the user to make the purchase. An assistive AI agent might want to provide interventions, such as reminding messages, to help humans achieve personal goals, such as reducing the amount of time spent on social networking sites.

If we assume the agent is *rational* and makes decisions according to the optimal policy, this problem is similar to several existing works in the literature, including policy teaching [209, 210], in which the principal updates the reward functions to induce the agent to take certain policies, and the poisoning attack for reinforcement learning [145, 212], in which an adversarial principal aims to modify the training environment such that the agent learns the undesired policy. In this work, we are motivated by the natural setting in which the agent is a human being and might exhibit biases in decision making. As observed in empirical studies, humans are known to exhibit systematic biases in making decisions. For example, humans might not have the ability to reason far ahead into the future [91] or might exhibit *present bias* [134], giving stronger weights on immediate costs and benefits rather than balancing them against those in the future.

We study this decision making problem under the formulation of MDP. Our setting deviates from the standard MDP in two perspectives. First, there are two parties, a principal and an agent, in the same decision making environment. The principal and the agent share the same information about the state, state transition, and action set. However, they have different reward functions. Moreover, while the agent can take actions in the environment, the principal can only update the environment to influence the agent’s actions. Second, the agent exhibits decision-making biases in his solution to the MDP. Since the focus of this paper is in sequential decision making, we focus on the *time-related decision biases*, including myopic decision making, bounded rationality, and present bias.

We consider two sets of design spaces that the principal can choose from to update the environment. In the first design space, the principal can modify the agent’s reward function in MDP, and the agent’s policy is based on the modified reward function. This design space corresponds to the scenario in which the principal can update the environment in a global manner (e.g., changing the badge design in social networking sites), and the agent will take actions in the updated environment. In the second design space, when the agent is choosing an action during decision time, the principal can offer additional incentives to nudge the agent to choose a different action. This design space corresponds to the scenario in which the principal can take interventions during the agent’s decision time (e.g., offering a coupon

when the user navigates to a certain page). In environment design with both design spaces, the goal of the principal is to maximize her own total rewards, depending on the principal’s reward function and the agent’s actions, subject to a budget constraint that the amount of environment updates is limited.

We formulate the principal’s environment design problems as constrained optimization problems under both design spaces. We first show that the optimization problems are generally NP-hard to solve for both design spaces. We then propose relaxed formulations and corresponding algorithms for solving the problems. To evaluate the effectiveness of our proposed algorithms for environment design, we conduct simulations to understand the algorithm performance over a range of scenarios and parameters. Moreover, to examine whether we can indeed update the environment to influence the decisions of real-world human decision makers, we conduct a human-subject experiment with 300 workers from Amazon Mechanical Turk. Our results demonstrate the environment updates derived by our algorithms can effectively influence humans’ decisions and lead to better total payoff.

3.1 Related Work

The work in this chapter is built on the formulation of Markov decision process commonly seen in reinforcement learning. Instead of solving the agent’s optimal policy, in this work, we consider a Stackelberg game formulation, in which the principal can first choose how to update the decision-making environment, and then the agent makes decisions in the updated environment. The goal of the principal is to obtain the maximum total rewards derived from the agent’s actions. When the agent is rational and chooses the optimal policy, our problem is similar to policy teaching [209, 208, 210] and poisoning attack for reinforcement learning [145, 212] in the literature. Our work deviates from these works by incorporating human behavioral models in the framework and in conducting real-world human subject experiments to evaluate our approaches.

We incorporate the human behavioral models about biased decision-making from behavioral economics. In particular, we include the *bounded rationality* [91], which describes the intuitions that human decisions might not be optimal due to limited computation power or lack of future information (a myopic agent can be considered as a bounded-rational agent that only cares about the current payoff), and *present bias* [134], which describes humans’

tendency to give stronger weights on immediate costs and benefits rather than balancing them against costs and benefits in the future. While these behavioral models are empirically observed to often better align with real human behavior, there are still relatively limited research that incorporate them in studying computational systems with humans in the loop.

There have been works that aim to incorporate behavioral models in computational frameworks. For example, the research on incentivizing exploration [63, 116] (see Chapter 11 in the recent survey by [164]) studies how a principal can incentivize myopic agents to perform exploration in bandit learning via designing specific monetary payments or information policies. Similarly, there have been studies that incorporate human models in bandit learning through accounting for the setting in which the arm quality is generated by humans that respond to algorithm behavior [110], the feedback of each arm pull is generated by humans with herding bias [171], and the reward distribution of each arm is influenced by the arm pull [174]. [172] relax the Bayesian rational assumption in incentive design; [100] and [101] study the planning for time-inconsistent agents in environments characterized by graphical models; [120, 121, 95] incorporate biased human model in goal recognition. Moreover, there have been works examining real-world human behavior in computational environments [211, 155, 73, 46, 47, 176]. Our work aligns with this line of research which incorporates realistic human behavioral models in computation. In particular, we focus on the study of environment design in sequential decision-making environments characterized by Markov decision process with biased decision makers.

There have been other lines of research that also includes humans in the loop of reinforcement learning frameworks. For example, inverse reinforcement learning [129, 1, 146] aims to infer the reward functions in MDP through observing demonstrations of the optimal policy. If the demonstrator is a human being, the demonstrations could be noisy or contain behavioral biases. There have been studies [58, 159, 87, 215] aiming to incorporate human behavioral biases in the inference process and infer both the rewards and biases simultaneously. While the research goal is different, this line of research complements our study in that the techniques can be applied to infer the reward function and human biases in our formulation. This work differs from this line of work in that our goal is to induce humans to perform desired behavior through finding optimal ways to update the decision-making environment instead of improving the learning algorithms.

3.2 Problem Formulations and Models

We discuss the problem formulation for environment design and two proposed methods in this section.

3.2.1 Environment Design Formulations

Decision-making environment. The decision-making environment formulation differs from the standard MDP introduced in Section 2.2. Specifically, we define the environment as a finite-time horizon MDP with two sets of reward functions: $W = \langle S, A, P, R^a, R^p, T \rangle$, where S is the set of states, A is the set of agent actions, $P(s'|s, a)$ is the transition probability from state s to state s' after taking action a , T is the time horizon, $R^a(s, a)$ is the bounded reward obtained by the agent after he takes action a at state s , and $R^p(s, a)$ is the bounded reward obtained by the principal after the agent takes a at state s .

Agent decision-making policy. Since the agent could be biased and might not make time-consistent decisions, we represent the agent policy in a time-inconsistent manner: $\Pi : S \times T \rightarrow A$. In particular, for an agent policy $\pi \in \Pi$, $\pi(s, t)$ denotes the action the agent will take in state s at time t when following policy π . We formulate the agent as a planner $H : W \rightarrow \Pi$, with input being an environment $w \in W$ and output being a policy $\pi \in \Pi$ according to his decision-making model. The agent’s goal is to maximize his *perceived* (potentially *biased*) rewards. To characterize the time-inconsistent behavior of the agent, we define the notion $d(t)$, the discounting factor that the agent perceives the payoff obtained t steps ahead. In the standard setting, $d(t)$ is often assumed to be in the form of γ^t with $\gamma \in (0, 1]$ being the time-discounting factor. In this paper, we address different forms of $d(t)$ that captures different agent models, which will be discussed later.

With $d(t)$ defined, we now characterize the agent policy by defining a *perceived* Q -function² $Q^\pi(s, a, t, \hat{t})$, specifying the agent’s perceived value at time t for him to take action a in state s at a future time $t + \hat{t}$ and follows policy π afterwards. This additional \hat{t} parameter captures the agent’s time-inconsistent belief: what the agent *thinks* he will do in a future time $t + \hat{t}$ while at time t might be different from what he will actually do at time $t + \hat{t}$. We also abuse the notation and let $\pi(s, t, \hat{t})$ denote the action the agent thinks what he would do in state

²This definition extends the standard Q -function to incorporate the agent’s biased decision making.

s in a future time $t + \hat{t}$ while at time t . This perceived $Q^\pi(s, a, t, \hat{t})$ can be expressed as the sum of (1) the perceived reward for taking action a in a future time step $t + \hat{t}$ while at time t : $d(\hat{t})R^a(s, a)$ and (2) the expected future reward for following policy π after $t + \hat{t}$: $\mathbb{E}[\sum_{t'=t+\hat{t}+1}^T d(t' - t)R^a(s_{t'}^\pi, \pi(s_{t'}^\pi, t, t' - t))]$, where $s_{t'}^\pi$ is the random variable denoting the state at t' if the agent follows π after $t + \hat{t}$. The expectation is over the randomness of the state transition.

Since the policy is only executed with $\hat{t} = 0$ ($\hat{t} > 0$ represents the agent's belief of what he would do \hat{t} steps ahead), we let $Q^\pi(s, a, t) = Q^\pi(s, a, t, 0)$ and $\pi(s, t) = \pi(s, t, 0)$. The agent policy π^* can then be written as:

$$\pi^*(s, t) = \operatorname{argmax}_a Q^{\pi^*}(s, a, t) \quad (3.1)$$

For a given environment, the agent policy can be solved by applying standard techniques, such as backward induction or value iterations as in Section 2.3.1.

Biased agent models. As discussed above, we use the notion $d(t)$, denoting how much the agent discounts the payoff t steps in the future to characterize the agent's behavior. This notion characterizes many common behavioral models, with some illustrative examples below:

- Standard model: in the literature, The agent is often assumed to have a consistent time-discounting factor $\gamma \in (0, 1]$ for discounting future payoff. Therefore, we can set $d(t) = \gamma^t$ to represent this standard assumption.
- Bounded rationality or short-sightedness: It considers the scenario in which the agent can only perform limited computation due to either time, cognitive, or information constraints. This can be approximated by considering that the agent only has information or only can reason about information within τ steps. We can formulate this by setting $d(t) = \gamma^t$ for all $0 \leq t \leq \tau$, and $d(t) = 0$ for all $\tau < t \leq T$. In the special case of *myopic agent*, who only cares about the immediate payoff and not the future payoffs, we can set $\tau = 0$.
- Present bias: When choosing between earning 10 dollars 100 days from now or 11 dollars 101 days from now, most people will choose the latter. However, when again being asked to choose between earning 10 dollars now or 11 dollars tomorrow, many people will change their decisions. This example illustrates the *present bias*, describing humans' inconsistency

in discounting future payoffs. One common way to account for this behavior is through hyperbolic discounting factor: $d(t) = \frac{1}{1+kt}$ for $k > 0$.

Design space of the principal. Recall that the principal aims to update the environment to influence the agent’s actions. We consider two natural sets of “updates” the principal can make to the environment:

- **Reward function modification:** The principal may pay costs to modify the agent’s reward function to influence the agent’s decisions. Formally, the principal can modify the agent’s reward from $R^a(s, a)$ to $\bar{R}^a(s, a) = R^a(s, a) + c(s, a)$ for taking action a in state s by paying a cost equal to the absolute value of the modification $|c(s, a)|$. The agent will only observe the modified reward function and will make decisions based on \bar{R}^a . Note that this type of environment updates is performed *offline* in the sense that it updates the environment before the agent starts to make their decisions in the environment.
- **Action nudge:** We also consider another design space, in which the principal can offer a non-negative incentive $c(s, a, t) \geq 0$ to *nudge* the agent to take action a in state s at time t . The agent’s reward in state s would then be $R(s, a) + c(s, a, t)$ if taking action a at time t while the future perceived rewards do not change. Different from the reward function modification, this nudge influences the agent’s decisions during *decision time*.

The principal’s goal is to maximize her total rewards derived from the agent’s actions under the budget constraint that the total cost does not exceed budget B . Given the agent’s policy π and the initial state distribution $p_0(s)$, let $p_t^\pi(s)$ be the state distribution at time t when the agent follows policy π , the principal’s total expected reward can be written as³:

$$\sum_{t=0}^T \sum_{s \in S} p_t^\pi(s) R^p(s, \pi(s, t)) \quad (3.2)$$

Hardness of environment design problem for biased agents. Before we discuss our proposed methods, we first present an important, although perhaps not surprising, result that if the agent exhibit biases in decision making, being oblivious of the biases could lead to

³We do not include the time-discounting factor for the principal’s payoff to simplify the presentations. Our results and discussion can be easily extended to the setting with time-discounting factor.

undesired outcome for the principal. The result showcases the importance of taking human behavior into account in environment design⁴.

Lemma 1. *If the principal performs environment design by assuming the agent is a standard agent while the agent is boundedly rational, the ratio between the principal’s reward after environment design compared with the principal’s reward obtained in environment design with the correct agent model could be arbitrarily close to 0.*

An intuitive example to prove lemma 1 is to consider a case where in order to receive a high reward, one must sacrifice his current reward. Myopic agent will never sacrifice, so the larger the high reward, worse myopic agent performed compared with optimal. More detailed discussion is available in the appendix.

3.2.2 Reward Function Modification

We first consider the environment design problem in which the principal can influence the agent’s decisions through modifying the agent’s reward functions $R^a(s, a)$. Let $c(s, a)$ be the modification the principal makes on $R^a(s, a)$, and $\bar{R}^a(s, a) = R^a(s, a) + c(s, a)$ is the reward function that the agent perceives and based on when making decisions. Let the updated MDP environment be \bar{w} , replacing the agent reward function as \bar{R}^a , and the agent policy on this environment be $\pi = h(\bar{w})$. The environment design problem for the principal is to choose the set of updates $\{c(s, a)\}$ to maximize her payoff subject to the budget constraint B . Again, let the initial state distribution be $p_0(s)$, and $p_t^\pi(s)$ be the state distribution at time t when the agent follows policy π , we can formulate the environment design problem as follows,

$$\begin{aligned} \max_c \quad & \sum_{t=0}^T \sum_{s \in S} p_t^\pi(s) R^p(s, \pi(s, t)) \\ \text{s.t.} \quad & \sum_{s \in S} \sum_{a \in A} |c(s, a)| \leq B ; \pi = h(\bar{w}) \end{aligned} \tag{3.3}$$

⁴All proofs are included in Appendix A.

Note that this is a bi-level optimization problem, in which the principal is optimizing over the space of $\{c(s, a)\}$ while the agent is optimizing his policy in response to the principal’s update in the form of $\pi = h(\bar{w})$. To solve the inner optimization problem (the agent’s optimal policy), we can define an updated \bar{Q}^π by replacing the reward R^a with \bar{R}^a and solve the policy π using backward induction. We show that this bi-level optimization problem is generally NP-hard to solve.

Theorem 2. *It is NP-hard to solve the environment design problem with reward function modification as defined in Equation (3.3).*

Relaxed formulation. To address this hardness result, we propose to use a soft-max stochastic policy ρ to relax the deterministic policy π . This relaxation makes the inner optimization differentiable, so first-order optimization methods might be applied. Instead of using $\pi(s, t)$ to denote the chosen action, we use $\rho(s, a, t)$ to represent the probability of choosing action a in state s at time t . Moreover, we again use \bar{Q}^ρ to denote the perceived cumulative reward for policy ρ . The definition is similar to Q^π except that we need to incorporate the randomness of policy when evaluating the future reward. Moreover, we use a soft-max form to approximate the agent policy: $\rho(s, a, t) = \frac{e^{\beta \bar{Q}^\rho(s, a, t)}}{\sum_{a'} e^{\beta \bar{Q}^\rho(s, a', t)}}$, $\forall s, a, t$.

Below we formulate the relaxed environment design problem. We now use $p_t^\rho(s)$ to denote the state distribution at time t (with $p_0^\rho(s)$ defined as the initial state distribution $p_0(s)$ for notational simplicity) when the agent follows policy ρ . In addition, we explicitly layout the state distribution over time following policy ρ as a constraint in the third constraint of the optimization problem. Since the gradient of the optimization variables exists, we can approach this optimization through a gradient-based algorithm, as in Algorithm 1.

$$\begin{aligned}
& \max_c \sum_{t=0}^T \sum_{s \in S} \sum_{a \in A} p_t^\rho(s) R^p(s, a) \rho(s, a, t) \\
& \text{s. t. } \sum_{s \in S} \sum_{a \in A} |c(s, a)| \leq B \\
& \rho(s, a, t) = \frac{e^{\beta \bar{Q}^\rho(s, a, t)}}{\sum_{a'} e^{\beta \bar{Q}^\rho(s, a', t)}}, \forall s, a, t \\
& p_{t+1}^\rho(s) = \sum_{s' \in S} \sum_{a \in A} p_t^\rho(s') P(s|s', a) \rho(s', a, t), \forall s, t \\
& \rho(s, a, t) \geq 0, \forall s, a, t
\end{aligned} \tag{3.4}$$

Algorithm 1 Gradient-based Algorithm for Solving Equation (3.4)

- 1: **Input:** learning rate δ , maximal iterations N
 - 2: initialize $c, i = 0$
 - 3: **while** $i < N$ **do**
 - 4: sample $\hat{s} \in S, \hat{a} \in A$
 - 5: update $\bar{R}^a(s, a), \bar{Q}(s, a, t), \rho(s, a, t), p_t^\rho(s), \forall s, a, t$
 - 6: calculate $\frac{\partial \rho(s, a, t)}{\partial c(\hat{s}, \hat{a})}, \frac{\partial p_t^\rho(s)}{\partial c(\hat{s}, \hat{a})}, \forall s, a, t$
 - 7: $c(\hat{s}, \hat{a}) \leftarrow c(\hat{s}, \hat{a}) + \delta \frac{\partial \sum p_t^\rho(s) R^p(s, a) \rho(s, a, t)}{\partial c(\hat{s}, \hat{a})}$
 - 8: $i \leftarrow i + 1$
 - 9: **end while**
 - 10: **return** c
-

Discussion. When $\beta \rightarrow \infty$, $\rho(s, a, t)$ approximates to a delta function with the probability mass on the action with the highest \bar{Q} value, which recovers the original problem. Moreover, recall that the Q function is defined with respect to the policy (when calculating the expected future rewards). We can show that this soft-max relaxation converges to the Q function of deterministic policy exponentially fast in β . In our simulations, we also empirically demonstrate that setting a small β is enough to approximate the optimal of the original problem in Equation (3.3).

Lemma 3. *For any environment w , let π_w and ρ_w be the agent's deterministic and stochastic policies following our model. Let $Q^{\pi_w}(s, a, t)$ and $Q^{\rho_w}(s, a, t)$ be the corresponding Q -functions. For all (s, a, t) , we have*

$$|Q^{\pi_w}(s, a, t) - Q^{\rho_w}(s, a, t)| \leq \mathcal{O}(e^{-\beta C}),$$

where $C > 0$ is a constant and β is the parameter of ρ .

3.2.3 Action Nudge

We now formulate the environment design problem via action nudge. The principal can choose to pay $c(s, a, t) \geq 0$ to the agent if he takes action a in state s at time t . In this approach, the agent's perceived Q function does not change, but the agent's action will be influenced by this additional incentive, i.e., the agent will choose the action that maximizes $Q^\pi(s, a, t) + c(s, a, t)$ in state s at time t . Moreover, since the nudge is calculated offline but deployed online, the budget constraint is satisfied in expectation. Formally, the principal's environment design problem can be written as:

$$\begin{aligned}
 & \max_c \sum_{t=0}^T \sum_{s \in S} p_t^\pi(s) R^p(s, \pi(s, t)) \\
 & \text{s.t.} \sum_{t=0}^T \sum_{s \in S} c(s, \pi(s, t), t) p_t^\pi(s) \leq B \\
 & \pi(s, t) = \operatorname{argmax}_a \{Q^\pi(s, a, t) + c(s, a, t)\}, \forall s, t
 \end{aligned} \tag{3.5}$$

Solving this problem directly is again generally NP-hard due to the same bi-level optimization property and the deterministic policy structure. Below we utilize the problem structure and develop an alternative formulation.

Alternative formulation. Let π be the agent's policy in the original decision-making environment. The goal of action nudge is to make the agent change from action $a = \pi(s, t)$ to a new action a' . Assume the principal can break ties in any way she prefers when multiple actions lead to the same payoff⁵, the cost the principal needs to pay to make the agent select action a' instead of a is $c(s, a', t) = Q(s, a, t) - Q(s, a', t)$. We can pre-calculate all the cost the principal needs to pay for action nudge $c(s, a, t) = Q(s, \pi(s, t), t) - Q(s, a, t), \forall s, a, t$.

⁵While this assumption seems strong, it can be approximately satisfied by adding an arbitrarily small value to $c(s, a', t)$ to make the agent break ties to align with the principal's goal.

With the above observations and the additional tie-breaking assumption, the environment design problem via action nudge is reduced to selecting which action the principal should nudge the agent to select for all (s, t) . The nudged action a would generate a reward of $R^p(s, a)$ and incurs a cost $c(s, a, t)$. The goal is to maximize the total rewards such that the total cost is no larger than budget B in expectation. This problem reduces to a standard constrained MDP problem.

$$\begin{aligned}
& \max_{\phi} \sum_{t=0}^T \sum_{s \in S} \sum_{a \in A} R^p(s, a) \phi(s, a, t) \\
& \text{s.t.} \sum_{t=0}^T \sum_{s \in S} \sum_{a \in A} c(s, a, t) \phi(s, a, t) \leq B \\
& \sum_{s' \in S} \sum_{a \in A} P(s|s', a) \phi(s', a, t) = \sum_{a \in A} \phi(s, a, t+1), \forall s, t \\
& \sum_{a \in A} \phi(s, a, 0) = p_0(s), \forall s \\
& \phi(s, a, t) \geq 0, \forall s, a, t
\end{aligned} \tag{3.6}$$

In this optimization problem, $\phi(s, a, t)$ is the optimization variables, representing the joint probability at time t for the agent to be in state s and take action a . To translate $\phi(s, a, t)$ to the stochastic policy $\rho(s, a, t)$, we have $\rho(s, a, t) = \frac{\phi(s, a, t)}{\sum_{a' \in A} \phi(s, a', t)}$. The optimization problem is a linear program in $\phi(s, a, t)$. Therefore we can directly apply standard linear programming solvers to solve this optimization problem. When the agent is in state s at time t , this solution indicates that the principal should nudge and offers $c(s, a, t)$ if $\phi(s, a, t) > 0$. There could be multiple actions that lead to $\phi(s, a, t) > 0$ for a given (s, t) , leading to offering multiple nudges simultaneously. Lemma 4 shows that there exists a solution such that this does not happen frequently and we can find such solution in polynomial time. The proof is available in Appendix A.4.

Lemma 4. *There exists an optimal solution ϕ^* for problem 3.6, such that there are at most one state-time (\hat{s}, \hat{t}) and two actions a_1, a_2 such that $\phi^*(\hat{s}, a_1, \hat{t}) > 0$ and $\phi^*(\hat{s}, a_2, \hat{t}) > 0$.*

3.3 Experiments

We conduct both simulated and real-human experiments to evaluate our proposed methods for environment design.

3.3.1 Simulations

In our simulations, we create a grid world of size 10×10 . Each grid represents a state in the MDP. There are four actions representing the direction agent can move to: {up, down, left, right}. After each action, the agent moves to the nearby grid associated with the action with 70% chance and to a random nearby grid with 30% chance. The initial state is in the middle of the grid world. The time horizon T is set to be 20.

We initialize the principal’s reward function values to be uniformly drawn from the range $[0, 0.5]$. We then randomly choose a 2×2 block as global optimal region and add 0.5 to the reward values within this block. Similarly, we randomly draw 1 to 3 local optimal regions (2×2 blocks) by setting their reward lower than global optimal but higher than its neighbors. We randomly generate 1,000 environments following the above procedure and report the average results. on these 1,000 environments.

Different agent behavioral models. We start with the setting that the agent’s reward function is the same as the principal’s, i.e., $R^p(s, a) = R^a(s, a)$ for all (s, a) . In this setting, if the agent is behaving optimally, the principal does not need to update the environment. Therefore, we focus on examining how the agent’s biased behavior impacts the total payoff and how effectively environment design can help.

We first examine the impact of biased agents without environment design. We consider agents with bounded rationality (or short-sightedness) and with present bias. Following the formulation in Section 3.2.1, we modify τ for boundedly-rational agents and k for present-bias agents. For boundedly-rational agents, we set $\gamma = 1$ and vary τ to be from 0 to 9. For present-bias agents, we vary k to be in $\{0.1, \sqrt{0.1}, 1, \sqrt{10}, 10\}$. The performance is measured in terms of the principal’s objective. As shown in Figure 3.1, the principal’s payoff, even when the reward function aligns with the agent’s, could decrease significantly when the agent exhibits decision biases.

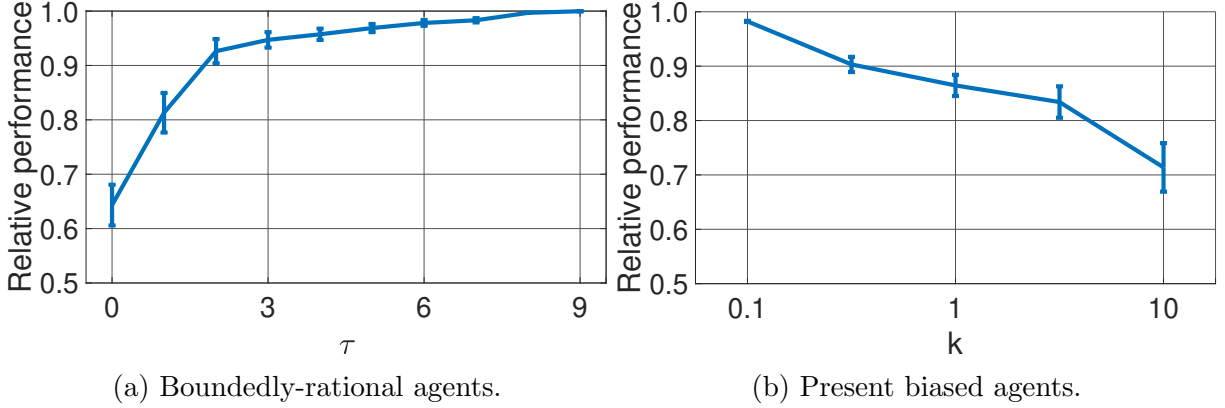


Figure 3.1: The principal’s payoff with biased decision-makers without environment design. Agents with higher τ or lower k are closer to being rational.

Next we examine the effect of environment design in improving the principal’s payoff. We apply the algorithms in Section 3.2.2 and 3.2.3, with the soft-max parameter $\beta = 3$ (the choice of β is discussed in the appendix). We examine present-bias agents with $k \in \{1, 10\}$ and boundedly-rational agents with $\tau \in \{0, 1, 2\}$. We vary the budget for algorithms with both design spaces. As in Figure 3.2, our algorithms lead to effective environment design and improve with larger budget. While action nudge seems more cost efficient, the cost needs to be incurred for each agent. In reward modification, the environment may need only be updated once for multiple agents.

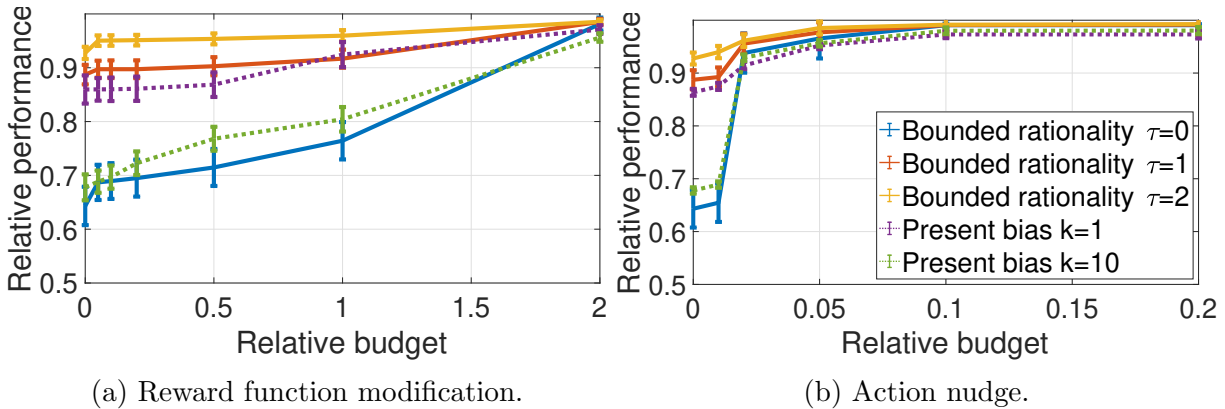


Figure 3.2: The principal’s payoff with biased decision-makers after applying environment design. The y-axis is the relative performance compared with the optimal solutions, and the x-axis is the amount of budget spent relative to the optimal performance.

Mis-alignment of the principal’s and the agent’s objective. We now consider the case that the agent’s reward function might not align with the principal’s. We fix the principal’s

reward function as before and vary the agent’s reward function. We consider the cases in which the agent’s reward function is the inverse (adversarial), randomly drawn (irrelevant), and the same (cooperative) of the principal’s reward function. The agent’s bias model is set to be boundedly rational with $\tau = 1$ (the results are qualitatively similar for other agent models). As shown in Figure 3.3, our algorithm can find the sets of environment updates to induce desired agent decisions, though it generally requires more budgets when the principal’s reward function does not align with the agent’s.

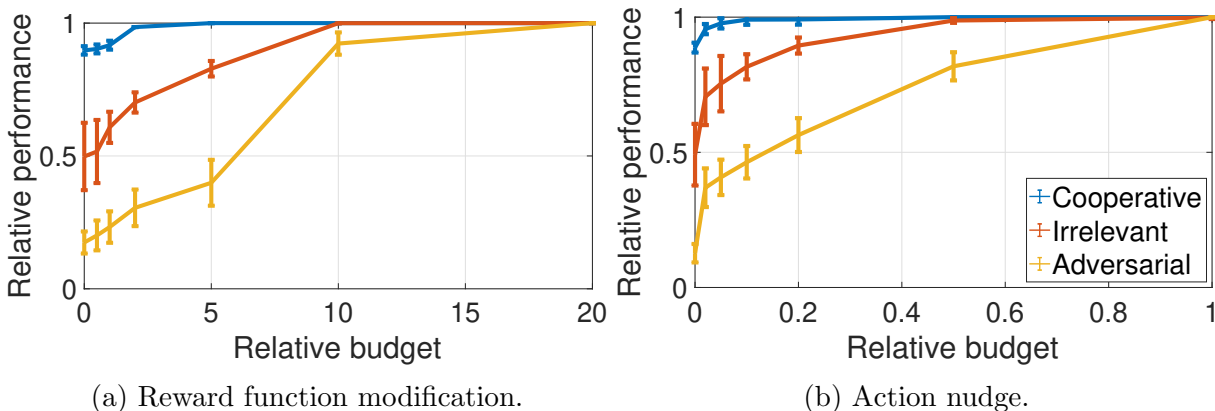


Figure 3.3: Misalignment of the principal’s and the agent’s the agent’s reward function. The y-axis is the relative performance compared with the optimal performance (in terms of the principal’s payoff), and the x-axis is the amount of budget spent relative to the optimal performance.

Additional simulations. Additional simulations are included in Appendix B.1. We show that setting a small β in Algorithm 1 suffices to approximate the true optimal of Equation (3.3) and examine its runtime. This result complements Lemma 3 and demonstrates that we can approximate the overall performance of the optimal. In another simulation, we demonstrate how to combine off-the-shelf inverse reinforcement learning algorithms to deal with scenarios when the agent rewards and biases are unknown a priori.

3.3.2 Human-Subject Experiments

While our simulation results are promising, they are under the assumption that the agent makes decisions following the behavioral model. In this section, we examine whether our environment design algorithms are effective for real human decision makers whose behavior

might deviate from the model. We have recruited 300 unique workers from Amazon Mechanical Turk. Each worker is paid \$0.50 and might earn additional bonuses. The average hourly rate is around \$11.50.

Task description. Each worker is asked to play six navigation games, with each represented by a grid world of size 10×10 . The setup is similar to our simulations, except that we simplify the rewards to depend only on the state, i.e., $R^a(s, a) = R^p(s, a) = R(s)$, to reduce the cognitive burden for workers. Workers’ bonuses depend on their total rewards. We also consider the setting in which the principal and the agent share the same reward function.

To induce biased human behavior, a worker can only see the rewards of the nearby states (to simulate the short-sightedness). Out of six games, there are two games each for vision length of 1, 2, 3, which we use short-sighted (boundedly rational) agent with $\tau = 0, 1, 2$ to model when solving the environment design problem. The detailed task setup is included in Appendix C.1.

Each worker is randomly assigned to one of the three treatments: {baseline, modified reward, action nudged}. The games are drawn from the same pool for each treatment. In baseline, workers play the drawn games without modifications. In modified reward, workers see the modified rewards generated by our algorithm. In action nudge, when a nudge happens, the workers see an additional messages indicating they might gain bonus for moving towards a certain direction. Since our goal is to observe whether environment design has impacts to real human decision-makers, we set the budget to be large enough such that the optimal decisions can be induced when the agent follows the behavioral model. We also report the true incurred cost in the experiment results.

Experiment results. As shown in Figure 3.4a, workers under both environment design treatments generate more rewards for the principal, suggesting that our algorithms lead to effective environment designs even for real humans that do not always behave as the behavioral model. The actual costs incurred in “modified reward” and “action nudge” treatments are 73.7 and 50.3 points, while the average gain is 142.9 and 119.2 points. Moreover, since the principal and the agent share the same reward, the baseline treatment corresponds to the optimal design (do nothing) for the standard agent model. The performance improvement of our algorithms re-affirms the importance of incorporating realistic human models.

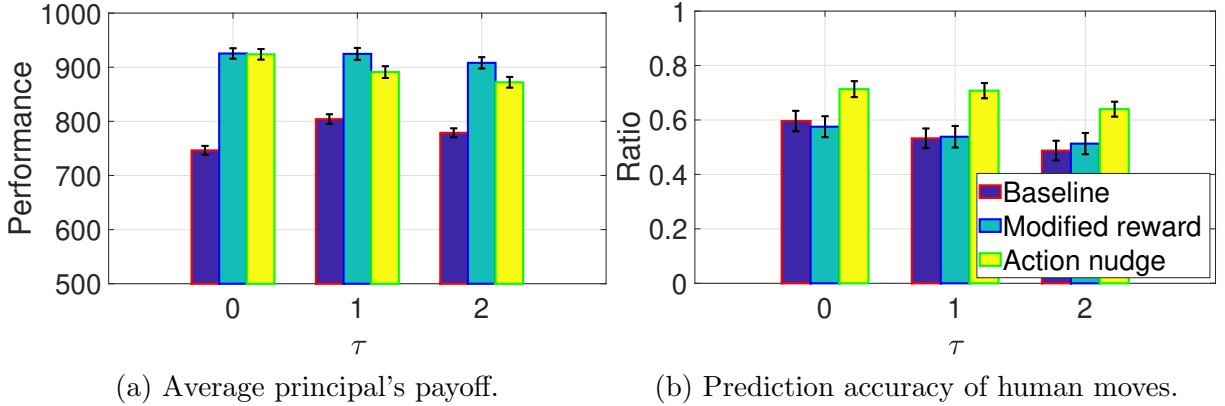


Figure 3.4: The human-subject experiment results of environment design. The results are grouped by the vision length of the games, mapping to different values of τ in short-sighted (boundedly rational) agents. Figure 3.4a shows the average principal’s payoff with real human decision makers in treatments, and Figure 3.4b shows the ratio of worker moves which are the same as short-sighted model predictions.

We also measure whether real humans behave as predicted by the behavioral model. As in Figure 3.4b, worker behavior aligns with our behavioral models 53.8%, 54.2%, 68.7% of the time on average in each treatment. We also compare worker behavior with the standard model, with alignment at only 33.2%, 36.9%, 45.9% of the time. Interestingly, workers are more likely to behave as predicted in the “action nudge” treatment, likely because this treatment generates additional information that triggers workers to follow the nudged action.

3.4 Discussions

We investigate environment design with biased decision makers. We explore two natural design spaces, reward function modifications and action nudges, and we formulate the environment design problems as constrained optimization problems under reward function modifications and action nudges. We first show neglecting biased decision maker leads to negative results. We then show that the environment design problems are NP-hard to solve and propose corresponding algorithms. We evaluate the algorithms through both simulations and human-subject experiments. Our work sheds lights on many important applications, such as AI-assisted decision making or adversarial machine learning.

Future works include incorporating other bias models, including different environment design strategies, and addressing potential concerns when the objective of the principal aligns with the agent, such as in settings that an AI agent aims to assist humans in making decisions, our framework provides insights on how AI can update the environment to better help humans. On the other hand, when the objectives of the principal and the agent differ, such as in the adversarial setting. For example, can we design robust decision-making environments, e.g., imposing regulations/constraints on the environment updates to be allowed, to better safeguard human welfare. Our framework helps understand the power of an adversary in updating the environment to sabotage human decisions. This understanding could help us investigate the design of robust decision-making environments, such as imposing regulations/constraints on the environment updates to be allowed, to better safeguard human welfare. We hope this work can open more discussion in designing assistive AI technology and in incorporating behavioral models in computation.

Chapter 4

Sending Persuasive Signals

In Chapter 3, we investigate how to modify decision-making environments to influence human decisions, using prior knowledge about time-inconsistent behavior to model real humans. However, our approach is limited by its focus on time-inconsistent behavior and specific types of biases. Human-subject experiment results reveal that real humans are possible to deviate from our (assumed) bias models. This raises the question of whether we can model real humans without relying on prior knowledge. In this chapter, we explore data-driven methods, using supervised learning to build human behavior models based on collected human responses. We use this data-driven method to describe human decision-makers and explore methods for presenting persuasive signals to influence human decisions. The work in this chapter is based on joint work with Wei Tang, Saumik Narayanan and Chien-Ju Ho [205]. I contributed to this work by conducting experiments to collect human responses, training human behavior models, proposing a general framework suitable for different receiver models, and running experiments to evaluate the proposed approach.

We utilize the information design framework to examine how AI systems can send information to human decision-makers to influence their decisions. The problem of information design is ubiquitous in various applications. For example, online retailers can highlight a subset of product features to influence buyers to make purchases [74, 115]. Recommendation systems might selectively display other users' ratings to persuade users to follow recommendations [185]. Politicians can influence voters' decisions by designing different policy experiments [5]. There have been various research efforts from economics [122, 147, 69, 71, 108], machine learning and artificial intelligence [198, 36, 7, 77, 25, 60], and general computer science [56, 48] devoted to the study of information design. Among the growing literature on information design, the model of Bayesian persuasion proposed by [94], is one of the most prominent, and has inspired a rich body of studies.

While Bayesian persuasion offers an elegant framework for formulating the information design problem, it has two limitations. First, the receiver is assumed to be Bayesian rational. This means that the receiver can form a posterior in a Bayesian manner and chooses the action that maximizes his expected utility.⁶ However, as consistently observed in empirical studies [124, 92], humans often deviate from being Bayesian or rational. Directly applying the techniques from the information design literature that assume Bayesian rational receivers could lead to suboptimal outcomes. In this work, we address this limitation by proposing a general optimization framework that can integrate a wide range of human behavior, expressed either as traditional analytical closed-form behavioral models or as data-driven models, and design optimal information policies with respect to the provided human behavior.

Second, despite a decade of effort, characterizing the optimal information policy remains notoriously difficult. [48] have shown that it is $\#P$ -hard to compute the optimal expected sender utility, and in multi-receiver settings where each receiver only has binary actions, it is $\#P$ -hard to even approximate the optimal sender utility within any constant multiplicative factor [197]. Moreover, most previous works have assumed that the receiver follows the Bayesian rational assumption. When this assumption is relaxed [38, 172], there are generally no known analytical solutions for finding the optimal information policy yet.

To address these two limitations, we encode human behavior into the design process. Inspired by the recent effort in utilizing deep learning for auction design [50, 135], we propose HAIDNet, an optimization framework that leverages neural-network architectures for information design. Unlike existing works that assume rational human behavior, our optimization framework can adjust to multiple representations of human behavior patterns, including standard behavioral models represented in analytic forms, and data-driven models trained using machine learning approaches. More specifically, we encode receiver behavior as a function and represent the loss in our optimization framework as a function of the receiver’s responses to the disclosed information. This approach enables our optimization framework to accommodate different representations of human behavior and can lead to corresponding optimal information policies. We then evaluate our approach via extensive simulations. We show that HAIDNet can recover the optimal information policies in simpler settings with known analytical solutions, and HAIDNet can extend to design information policies for settings that are computationally challenging (e.g., multiple receivers involved), or for settings with no

⁶We use she/he to denote the sender/receiver respectively.

known solutions in general (e.g., when the receiver’s behavior does not follow the standard Bayesian rationality assumption).

4.1 Related Work

The work in this chapter joins a growing number of studies that leverage computational tools for automated mechanism design [34, 152], the problem of utilizing computational approaches or learning-based techniques for finding revenue-maximizing mechanisms in auction settings. One strand of works [33, 153] in this line of research has focused on using learning approaches for mechanism design where only samples of bidder valuations are used to design revenue-maximizing mechanisms. More recently, deep neural networks has been utilized for the automated design of optimal auctions [50], in which the authors propose multiple neural-network architectures for learning approximately optimal auctions. Several works has extended this study in various applications [61, 72, 37, 103, 144, 135, 109, 36]. Our work differs from this line of works in two ways. First, we extend the approach beyond auction design to address the automated information design problem. Second and more importantly, we have incorporated human behavior descriptors in our design, while prior works mostly require standard rationality assumptions.

Our information design formulation builds on top of the seminal work of Bayesian persuasion [94], which initiated a rich theoretical literature on communication games in which a sender can design information to persuade a receiver to take certain actions. Their work has provided theoretical foundations and inspired an active line of research in information design (e.g., see the recent surveys by [93, 18]). Our work builds on top of this line of work through integrating human behavior in the design of information policy, while existing works mostly assume the receiver is Bayesian rational. In particular, our proposed HAIDNet can dynamically adjust to various forms of model-based or data-driven human behavior descriptors. For the model-based receiver behavior, as an example, we have included the probability weighting function [195, 141, 149] for belief updating and the discrete choice model [123, 165, 179] for decision making under uncertainty. Non-Bayesian belief updating in information design also appears in earlier works [38], and the receiver’s behavior following the discrete choice model also appears in previous works [172, 59]. Our work generalizes the above in that our framework can adapt to both the above form and the data-driven form of human behavior.

The problem of information design and persuasion has received increasing attention both in research and in practice. For example, researchers have argued that one-quarter of the GDP in the United States is persuasion [122]. Due to its practical relevance, this problem is getting attention more broadly in the general research community, as demonstrated by the recent papers in machine learning and artificial intelligence venues, studying various problem settings such as in security [198], human language interactions [7], data marketplace design [28], algorithmic recourse [77], online recommendation [60], and market competitions [42]. Our work joins this line of study and aims to develop more efficient approaches for information design under more realistic settings of human behavior.

On a conceptual level, this work is related to the growing attention in understanding, modeling, and accounting for human behavior in computational systems, especially in the context of human-robot or human-AI interactions [31, 159, 104, 26, 148, 128, 127, 180, 181]. Moreover, our work joins the recent research theme that incorporates human models in computational and machine learning frameworks [63, 116, 171, 172, 100, 120, 121, 175, 203]. There have been other lines of research that includes humans in the loop of learning frameworks, such as inverse reinforcement learning [129, 58, 159, 87, 215] that infers the reward functions in Markov decision process through (potentially human) demonstrations. Our work differs in that we focused on the information design problem with realistic human receiver models.

Lastly, in this study, we incorporate insights from human behavior into information design. Extensive literature from psychology and behavioral economics has been devoted to deepen our understanding of human behavior. Examples include studies examining deviations from the standard Bayesian assumption in processing information [130, 97, 13] and the rationality assumption in decision-making [92, 123, 165, 179, 81]. While these classical models, often grounded in human data from behavioral experiments [118, 46, 47], offer interpretable behavioral insights, they tend to lack in terms of predictive accuracy. Recently, given the advancements of machine learning techniques and the availability of a larger amount of human data, there has been a growing effort to integrate behavioral insights from these classical models with machine learning techniques to enhance predictive accuracy [22, 138]. These models developed in this line of effort are directly applicable in our framework. Moreover, integrating human behavioral insights into information design can raise concerns about exploiting human irrationality. One potential solution is to incorporate the concept of differential privacy [52, 51, 173]. This would control the amount of personalized information that can be used, preventing undue exploitation.

4.2 Problem Formulations and Models

We introduce the problem formulation of Bayesian Persuasion and our proposed AI model (the sender) in this section.

4.2.1 Bayesian Persuasion Formulations

In Bayesian persuasion, there are two players: a sender and a receiver. The sender's goal is to design an information disclosure policy that persuades the receiver to take certain actions maximizing the sender's objective. The state of nature θ is drawn from a finite set $\Theta \triangleq \{1, \dots, m\}$ according to a prior distribution $\lambda \triangleq (\lambda(\theta))_{\theta \in \Theta} \in \Delta(\Theta)$. The prior is common knowledge to both the sender and the receiver. The receiver's utility $u^R(a, \theta)$ depends on the receiver action $a \in \mathcal{A}$ from an action set \mathcal{A} and the state θ . The sender's utility $u^S(a, \theta)$ also depends on the receiver's action and the state.

The sender can observe the realized state while the receiver cannot, and the sender can utilize this information advantage to persuade the receiver to take the desired action. In particular, before observing the realized state, the sender can commit to an information policy π , specifying what signal to present to the receiver conditional on the realized state. More formally, an information policy π consists of a signal space Σ and a set of conditional probabilities $\{\pi(\cdot|\theta)\}_{\theta \in \Theta}$ where $\pi(\cdot|\theta) = (\pi(\sigma|\theta))_{\sigma \in \Sigma} \in \Delta(\Sigma)$ and $\pi(\sigma|\theta) \in [0, 1]$ denotes the probability to send signal $\sigma \in \Sigma$ given the realized state θ . This information disclosure policy is known to the receiver and specifies how the sender discloses information to the receiver. When a state $\theta \in \Theta$ is realized, the sender sends a signal $\sigma \sim \pi(\cdot|\theta)$ according to the policy.

In Bayesian persuasion, the receiver is assumed to be Bayesian rational in the sense that upon seeing the signal σ , the receiver forms his posterior belief about the state in a Bayesian manner and takes an action that maximizes his expected utility. Formally, upon seeing the signal realization σ , the receiver updates his posterior belief over the state of nature, denoted by $\mu(\sigma) \triangleq (\mu(\theta|\sigma))_{\theta \in \Theta} \in \Delta(\Theta)$, by applying Bayes' rule: $\mu(\theta|\sigma) \triangleq \frac{\pi(\sigma|\theta)\lambda(\theta)}{\sum_{\theta' \in \Theta} \pi(\sigma|\theta')\lambda(\theta')}$.

Given the posterior induced from the observed signal $\sigma \in \Sigma$, the receiver takes an action $a^{\text{BR}}(\sigma) \in \mathcal{A}$ that maximizes his expected utility⁷, namely, $a^{\text{BR}}(\sigma) \triangleq \operatorname{argmax}_{a \in \mathcal{A}} \sum_{\theta \in \Theta} \mu(\theta|\sigma) u^R(a, \theta)$.

⁷BR here stands for Bayesian rational.

The sender’s information design problem is to find the optimal information policy that maximizes her expected payoff induced by the receiver’s action, as follows:

$$\max_{\pi} \sum_{\theta \in \Theta} \lambda(\theta) \sum_{\sigma \in \Sigma} \pi(\sigma|\theta) u^S(a^{\text{BR}}(\sigma), \theta). \quad (4.1)$$

In this work, our goal is to design an automated framework to solve the above bi-level optimization problem while encoding realistic human behavior in the design process (i.e., replacing the Bayesian rational human model $a^{\text{BR}}(\sigma)$ with general human behavior).

Example. Consider the scenario in which an online retailer (the sender) aims to persuade a buyer (the receiver) to make a purchase. The retailer’s products are directly coming from the factory, and the product quality (represented by the binary state θ) is drawn from a prior distribution λ . The buyer’s utility $u^R(a, \theta)$ depends on both his binary purchase decision a and the binary product quality θ , while the retailer’s utility $u^S(a, \theta) \equiv u^S(a), \forall \theta$ is state-independent and only depends on the buyer’s purchase decision. For example, the goal of the retailer is to persuade the buyer to make a purchase, i.e., $u^S(a) = 1$ for $a = 1$ and $u^S(a) = 0$ for $a = 0$. The buyer only wants to purchase when the product is good, i.e., $u^R(a, \theta) = 1$ if $\theta = a$, and $u^R(a, \theta) = 0$ otherwise.

In order to persuade the buyer to purchase, the retailer can commit to performing (noisy) product inspections $\pi(\sigma|\theta)$ to reveal information about the product quality. For example, the inspection might signal the product quality is satisfactory with 80% chance if the quality of the product is indeed satisfactory (i.e., $\pi(\sigma = 1|\theta = 1) = 0.8$) and signal the product quality is unsatisfactory with 90% chance if the quality is indeed unsatisfactory (i.e., $\pi(\sigma = 0|\theta = 0) = 0.9$). The information design problem for the retailer is to identify an inspection policy that maximizes the probability on selling the product to the buyer.

4.2.2 Encoding Human Behavior in Information Design

To solve the problem, we introduce HAIDNet, an optimization framework based on a neural network architecture that can adjust to various forms of human behavior. In the following discussion, we first describe how we modularize human behavior in information design. We then explain the neural network architecture of our proposed HAIDNet that can adapt to

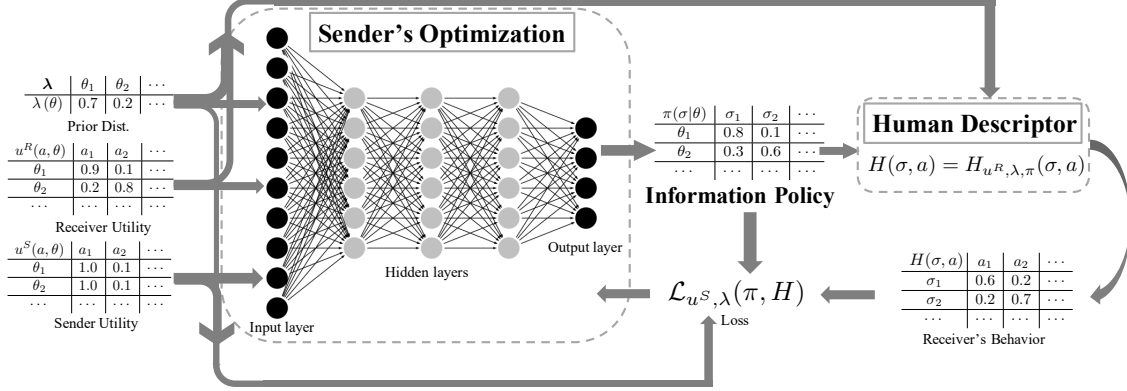


Figure 4.1: The HAIDNet framework. The human descriptor module is given to the optimization module before training. The optimization is performed through back propagation which evaluates the gradient of the loss to update the weights in the neural network structure.

different forms of human behavior. Finally, we outline the procedures for optimizing the information policy in HAIDNet.

Bayesian persuasion assumes that the receiver is Bayesian rational. However, in practice, this assumption often does not hold. The receiver may exhibit systematic biases both in belief updating and in decision making. In the following discussion, we formulate the sender's problem on finding the optimal information policy when taking more general human behavior into account.

Human Behavior Descriptor. For any receiver utility u^R , prior λ , sender information policy π (and signal space Σ), a human behavior descriptor is denoted by $H_{u^R, \lambda, \pi}(\sigma, a)$, representing the probability for a human receiver to choose action $a \in \mathcal{A}$ when seeing a realized signal $\sigma \in \Sigma$.

When the context is clear, we omit the subscripts and write $H_{u^R, \lambda, \pi}(\sigma, a)$ as $H(\sigma, a)$ for notational simplicity. With the above definition, we can rewrite the sender's information design problem as:

$$\max_{\pi} \sum_{\theta \in \Theta} \lambda(\theta) \sum_{\sigma \in \Sigma} \pi(\sigma|\theta) \sum_{a \in \mathcal{A}} H(\sigma, a) u^S(a, \theta). \quad (4.2)$$

Below we give a few examples of human behavior descriptors.

Bayesian rational (BR). In standard Bayesian persuasion, the receiver updates his posterior in a Bayesian manner and takes action that maximizes the expected utility. Following the definition in Section 4.2.1, the human descriptor can be written as $H(\sigma, a) = \mathbf{1}\{a = a^{\text{BR}}(\sigma)\}$.

Probability weighting and discrete choice (TH-Model) We present another human behavior descriptor based on the work by [172] (denoted as the TH-model in the description of this chapter). In particular, they combine probability weighting, assuming the receiver’s posterior is distorted based on a function $\omega(\cdot) : \Delta(\Theta) \rightarrow \Delta(\Theta)$, and discrete choice model, assuming the receiver’s action is stochastic, with a higher probability in taking an action with higher expected utility (based on the distorted posterior belief).

Formally, let $\omega(\theta|\sigma)$ be the receiver’s distorted posterior belief after seeing signal σ and β_H be a parameter in the discrete choice model that tunes how stochastic the receiver’s action is (when $\beta_H \rightarrow \infty$, the discrete choice model reduces to standard expected utility maximization), the human behavior descriptor for this model can be written as:

$$H(\sigma, a) = \frac{\exp(\beta_H \sum_{\theta \in \Theta} \omega(\theta|\sigma) u^R(a, \theta))}{\sum_{a'} \exp(\beta_H \sum_{\theta \in \Theta} \omega(\theta|\sigma) u^R(a', \theta))}. \quad (4.3)$$

Data-driven human behavior descriptor. Note that in our formulation, we use the function $H(\sigma, a)$ to represent human behavior. Suppose we have access to sufficient human behavioral data, instead of expressing $H(\sigma, a)$ using a closed-form analytical expression, we can train a machine learning model to approximate this function and utilize the learned model as the human behavior descriptor.

We now introduce the framework of HAIDNet and explain how we utilize it to optimize the sender’s information policy for a given human descriptor.

HAIDNet framework. As presented in Figure 4.1, HAIDNet consists of two modules: the sender’s optimization module and the human descriptor. The sender’s optimization module is a neural network responsible for optimizing the sender’s optimal information policy. It takes the information design problem instances as input, including the prior distribution λ over the states and the payoff functions u^S, u^R for all players. The module outputs an

information policy which consists of a set of conditional probabilities $\{\pi(\cdot|\theta)\}_{\theta \in \Theta}$ over the signal space for each state $\theta \in \Theta$.

The human descriptor can either be model-based (e.g., Bayesian rational model or TH model in Equation (4.3)), or data-driven (e.g., a neural network modeling the receiver’s behavior). The human descriptor is treated as a black box from the perspective of the sender’s optimization module, and is fixed before HAIDNet begins training. The input of the descriptor consists of the receiver utility u^R , the prior distribution λ , and the information policy π (i.e. the output of the sender’s optimization module), while the output is the receiver’s response strategy $H(\sigma, a) = H_{u^R, \lambda, \pi}(\sigma, a)$.

Optimization procedure. For the sender’s optimization, we follow the recent line of research on using deep learning for auction design [50]: we randomly draw problem instances from a pre-specified distribution and perform stochastic gradient descent to minimize the loss function in the training process. The loss function is defined to be the negative of the sender’s expected utility, since the goal of the sender is to find the optimal information policy that maximizes her expected utility.

$$\mathcal{L}_{u^S, \lambda}(\pi, H) = - \sum_{\theta \in \Theta} \lambda(\theta) \sum_{\sigma \in \Sigma} \pi(\sigma|\theta) \sum_{a \in \mathcal{A}} H(\sigma, a) u^S(a, \theta) . \quad (4.4)$$

Our work differs from previous works in that we incorporate the human behavior descriptor in the definition of the loss function. The requirement is that the human descriptor $H(\sigma, a)$ needs to be differentiable. This requirement is naturally satisfied in many cases, e.g., when the human descriptor follows the model defined in Equation (4.3) or is a neural-network-based model, the gradient always exists. However, in the Bayesian rational model, since the receiver chooses the action that maximizes his expected utility, this *argmax* operation makes the human model not differentiable. To overcome this issue, we approximate the Bayesian rational model using softmax instead of argmax with a sufficiently large softmax scale parameter β .⁸ More concretely, let $u(a)$ be the expected utility for action a . The softmax operator approximates the receiver’s behavior by using $\exp(\beta u(a)) / \sum_{a'} \exp(\beta u(a'))$

⁸The notation β here is different from β_H used to model human behavior in Equation (4.3).

to denote the probability of choosing action a . As a sanity check, when $\beta \rightarrow \infty$, this expression reduces to argmax , choosing the action maximizing the expected utility.

To optimize HAIDNet, we train a neural network with 3 fully connected layers employing ReLU activation functions and the Adam optimizer. The model is trained on 100 batches of size 1024, for a total of 102,400 uniformly drawn problem instances (i.e., data points for training). Evaluation of the model is conducted on a test set consisting of 1000 problem instances. The specification of hyperparameters and implementation details are included in the Appendix B.2.

4.3 Experiments

We conduct both simulations and human-subject experiments to evaluate proposed HAIDNet.

4.3.1 Simulations

Our simulation results demonstrate that HAIDNet can find the near-optimal information policy in various settings. Specifically, we show its effectiveness in settings where efficient methods exist to obtain the optimal information policy and in computationally challenging settings where finding the optimal information policy is difficult. Moreover, even in settings where no known solutions exist in general, HAIDNet can generate information policy with good performance.

We have conducted additional simulations, including examining the convergence of the training, investigating the scalability of the approach, accounting for varying number of receivers, comparing with random policy, and examining empirical run-time. Additional simulation results are included in Appendix B.2.

We start our evaluations with a simple setting where there exist efficient solutions to find the optimal policy. In this setting, we leverage the efficient solutions as ground truth to examine whether our approach can also identify the optimal information policy.

In particular, we consider the setting with a single Bayesian rational receiver. In this setting, when there are only two actions available for the receiver and there are only two states, there exists a closed-form characterization of the optimal information policy. When the numbers of actions and states are finite constants, the optimal information policy can still be computed efficiently [94]. Therefore, we can evaluate the performance of our approach by comparing the information policy generated by HAIDNet with the optimal policy.

Binary actions and binary states. We first examine the simplest setting with binary actions and binary states (a classical setting in Bayesian persuasion [94]), namely, the action space $\mathcal{A} = \{0, 1\}$ and the state space $\Theta = \{0, 1\}$, and observe whether HAIDNet produces near-optimal information policies. For the sender utility, we adopt a stylized setting where the sender obtains utility 1 when the receiver takes action 1 and utility 0 when the receiver takes action 0. The receiver aims to take the action that aligns with the true state, i.e., $u^R(0, 1) = u^R(1, 0) = 0$, and we randomly draw each value for $u^R(0, 0)$ and $u^R(1, 1)$ from $[0, 1]$. In plain words, the receiver prefers action 1 when the state is 1 and action 0 when the state is 0, and the goal of the sender is to persuade the receiver to take action 1. The prior distribution λ is drawn from a Dirichlet distribution. We then simulate data using the setting above and optimize HAIDNet.

We first examine whether the policy generated by HAIDNet matches the known optimal policy. Note that in this simple setting, via revelation principle [94], an information policy can be characterized by two signals, i.e., $\sigma \in \{0, 1\}$, where each signal corresponds to a recommended action. Moreover, in the optimal policy, we have $\pi^*(\sigma = 1|\theta = 1) = 1$, and therefore the optimal policy can be characterized by a single parameter $\pi^*(\sigma = 1|\theta = 0)$. To examine whether HAIDNet generates the same policy as the optimal policy, we compare the value of this parameter on different scenarios.

To showcase our results, we present two settings where we have fixed prior distributions: low prior with $\lambda(\theta = 0) = 0.3$ and medium prior with $\lambda(\theta = 0) = 0.5$.⁹ For each prior distribution, we vary the receiver utilities and report the parameter $\pi^*(\sigma = 1|\theta = 0)$ both from the optimal policy and from the output of HAIDNet. As visualized in Figure 4.2, the policy learned by HAIDNet essentially recovers the optimal information policy in almost all scenarios.

⁹The results are the same for a wide range of prior distributions.

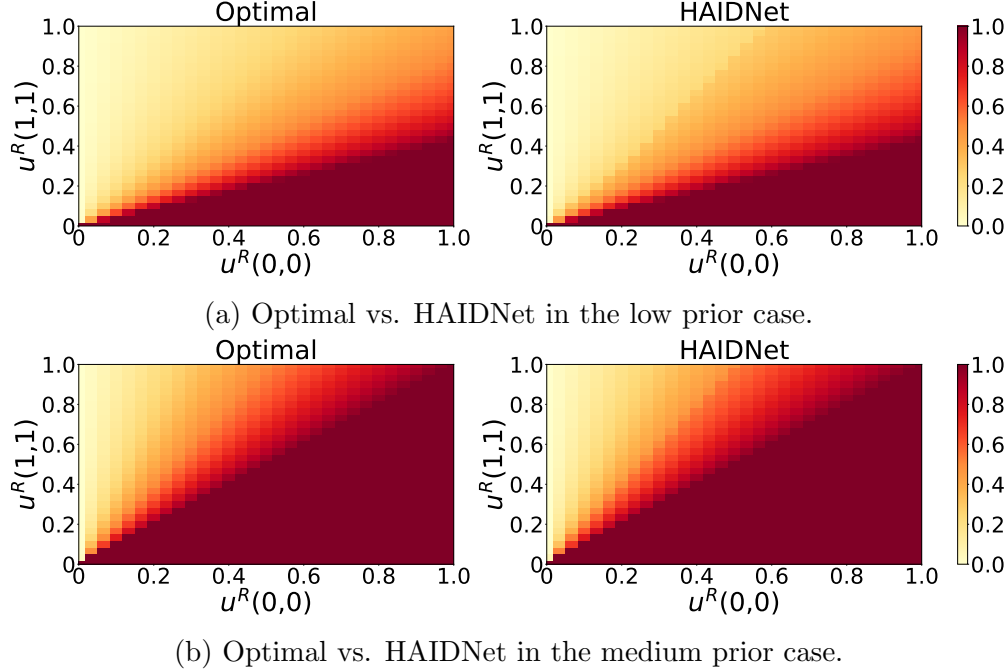


Figure 4.2: Comparing the optimal information policy and the policy generated by HAIDNet in the setting with binary actions and binary states.

Multiple actions and multiple states. To examine whether our approach scales with the size of the problem instances, we increase the number of states and the number of actions¹⁰. The performance is measured using the average sender utility. We report both the training performance (e.g., average sender utility for 1,000 instances drawn from instances used for training HAIDNet) and testing performance (e.g., average sender utility for newly drawn 1,000 instances).¹¹ The results, as shown in Table 4.1, demonstrate that our approach works well for large-scale problem instances and also generalizes well to instances not used in training.

Next, we examine the performance of HAIDNet under the setting where there are no known computationally efficient solutions to characterize the optimal information policy. The goal is to illustrate that HAIDNet performs well even in complicated scenarios and could provide a more efficient approach for settings without analytically tractable solutions.

¹⁰The results for scaling up both simultaneously are qualitatively the same and are included in the appendix.

¹¹We have included additional comparisons to the performance of a simple baseline, random policy, in the appendix. The performance for the random policy is around 0.5 in all scenarios in this setting.

Table 4.1: Comparing the average sender utility generated by the optimal policy and the policy from HAIDNet in the setting with a single Bayesian rational receiver.

(a) Increase the number of states M with binary actions. (b) Increase the number of actions N with binary states.

M	Training		Testing		N	Training		Testing	
	HAIDNet	Optimal	HAIDNet	Optimal		HAIDNet	Optimal	HAIDNet	Optimal
2	0.7409	0.7498	0.7408	0.7451	2	0.7409	0.7498	0.7408	0.7451
3	0.7737	0.7782	0.7598	0.7669	3	0.7017	0.7214	0.7089	0.7227
5	0.8171	0.8209	0.8066	0.8225	5	0.6906	0.7113	0.6690	0.7064
10	0.8495	0.8699	0.8196	0.8686	10	0.6861	0.7084	0.6623	0.6963

We consider the setting with multiple receivers and binary actions. The goal is to design a uniform information policy for all receivers (i.e., *public persuasion* [197]). This setting has been shown to be $\#\mathcal{P}$ -hard to find a policy that approximates the optimal sender utility within any constant multiplicative factor [49]. This means that, unlike the single receiver case, finding the optimal solution for a given problem is practically impossible to solve with a large set of receivers, and we intend HAIDNet to be a new, efficient solver for near-optimal solutions. To examine whether HAIDNet finds the optimal policy, we utilize a brute-force linear-programming approach [49] (the time complexity is exponential in the number of receivers since the number of constraints in the program grows exponentially) to identify the optimal policy when the number of receivers is small. We then compare the information policy generated by HAIDNet and the optimal policy output from the linear programming approach. The receiver utility and prior distributions are generated in the same way as in the single receiver setting. The sender utility is the fraction of receivers choosing action 1, i.e., her utility is given $\frac{|S|}{K}$ if there are $|S|$ receivers choosing action 1 out of a total K receivers.

The simulation results are shown in Table 4.2. We randomly draw 1,000 problem instances from the training/testing set and report the average performance of the optimal policy and the HAIDNet policy. As we can see in the results, the performance of the information policy output from HAIDNet is near-optimal. Moreover, HAIDNet provides a much more efficient approach when the number of receivers is large. As a comparison, solving the exact optimal information policy for each problem instance is time-consuming (e.g., it takes more than 23 hours to solve an instance with 18 receivers). On the other hand, HAIDNet only needs to optimize the model once to generate the optimal information policies for all possible problem instances with the same number of receivers (e.g., training HAIDNet with 18 receivers takes

slightly more than 1 hour, and generating information policy for a problem instance takes less than 1 second). The empirical run-time comparison is included in the appendix.

Table 4.2: Comparing the average sender utility generated by the optimal policy and the policy from HAIDNet in the setting with K Bayesian rational receivers.

K	Training		Testing	
	HAIDNet	Optimal	HAIDNet	Optimal
2	0.7887	0.7934	0.7756	0.7873
3	0.7508	0.7665	0.7379	0.7573
5	0.7217	0.7458	0.7209	0.7570
10	0.6971	0.7152	0.6790	0.6966
15	0.6553	0.6882	0.6621	0.6843

Non-Bayesian-rational receiver case. We now examine the performance of HAIDNet in settings where there are generally no known analytical solutions yet. The goal is to showcase that HAIDNet can be leveraged to address information design problems when we do not have access to solutions.

All our simulations so far have focused on settings which assume that receivers are Bayesian rational. To examine whether HAIDNet works for non-Bayesian-rational receivers, we adopt a relaxation of human behavioral formulation as in Equation (4.3). While there are no known solutions for identifying the optimal policy in this setting in general, [172] derived a solution for the simple setting with binary actions and binary states. Therefore, we compare the performance of the optimal policy and the HAIDNet policy in this simple setting under different choices of β_H in the human descriptor in Equation (4.3). Using the same setup as in previous simulations, we report the results in Table 4.3, showing that HAIDNet works even for a non-Bayesian-rational receiver.

Next, we would like to examine how HAIDNet performs in scenarios when there are no known solutions (e.g., in settings with more than binary actions/states). To demonstrate the results, we choose the setting with three states and three actions. The lack of an optimal solution means we cannot evaluate the performance of HAIDNet by comparing its performance with the optimal policy as in the simulations above. Instead, we take a different method and provide evidence to support our approach: We evaluate the set of all learned policies π_{β_H} against each of the human models β_H .

For each human model $\beta_H = k$, if π_k is the best-performing policy, this indicates that our approach generates a reasonably good information policy. Specifically, for each human

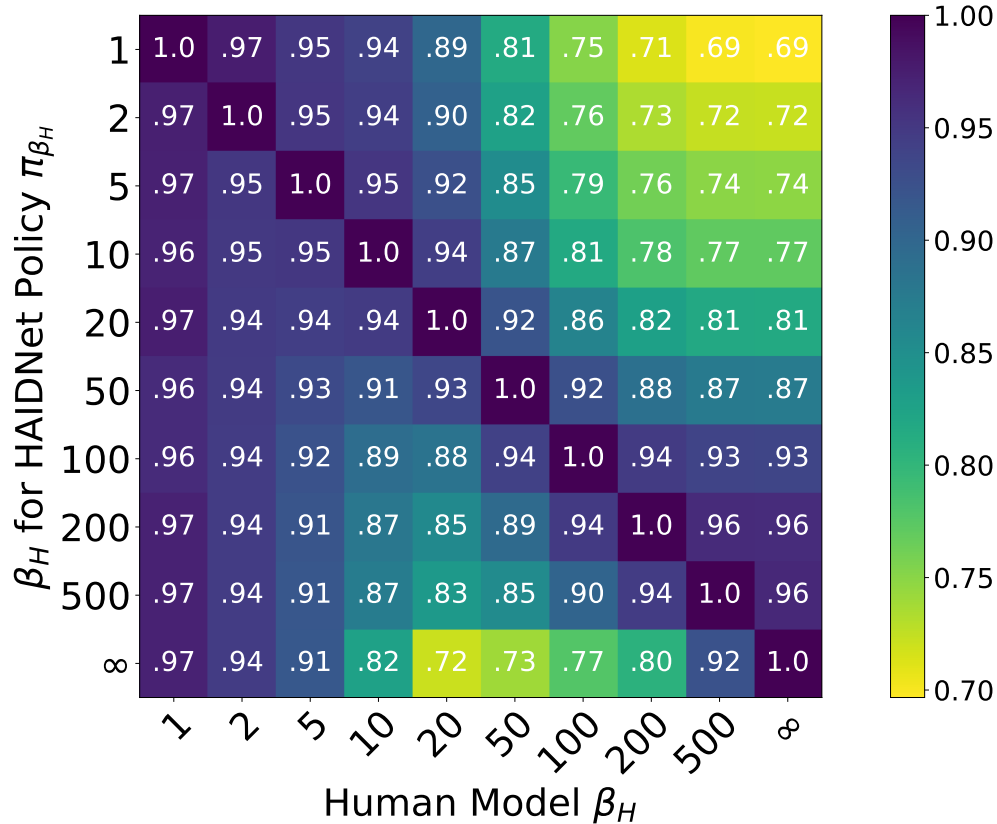


Figure 4.3: The performance of HAIIDNet in settings when the receiver is not Bayesian rational. We train HAIIDNet with non-Bayesian-rational receiver model parameterized by β_H , then evaluate the learned information policy for all receiver models. The performance is normalized so for each human model, the optimal performance is 1.0 among all policies.

model, we compute the performance of each policy available, and we then normalize the set of these performances so that the best-performing performance for each human model has value 1. If our HAIIDNet indeed learns a good information policy, we would expect the best performing HAIIDNet to be the one trained on the right human descriptor. The results, as shown in Figure 4.3, demonstrate this behavior and provides evidence that our HAIIDNet generates good information policy even when the receiver is not Bayesian rational.

4.3.2 Human-Subject Experiments

In the simulations, we have assumed access to a closed-form behavior model of the receiver. However, in practice, human behavior is complex and there may not exist a single model

Table 4.3: Comparing the average sender utility by the optimal policy and the policy from HAIDNet in the setting with a non-Bayesian-rational receiver parameterized by β_H .

β_H	Training		Testing	
	HAIDNet	Optimal	HAIDNet	Optimal
1	0.5043	0.5051	0.5041	0.5060
5	0.5512	0.5557	0.5506	0.5559
10	0.6045	0.6170	0.5986	0.6168
50	0.7002	0.7134	0.6800	0.7081
100	0.7187	0.7291	0.6964	0.7179

that can perfectly represent human behavior. Motivated by this practical concern, we conduct human-subject experiments to examine whether HAIDNet adapts to real-world human behavior. The goal is to examine whether we can utilize data-driven approaches to learn human-behavior descriptors and examine whether HAIDNet performs well when it is paired with data-driven behavior descriptors.

Task description. In our human-subject experiments, we present the product purchasing example in Section 4.2.1 to human participants. Each human participant is asked to make multiple rounds of purchase decisions. In each round, the participant is presented a product with unknown binary quality (good or bad product). The participant is told that a (noisy) inspection has been performed on the product, and is given the conditional distribution associated with the inspection (i.e., the probability to receive a good/bad signal given the product is good/bad). Finally, the participant is given a realization of the inspection signal and is asked to make a binary decision of purchasing or not. The participant’s payment depends on both their purchasing decisions and the true product quality. The task interface is included in Appendix C.2. The experiment is approved by the IRB in Washu.

Experiment procedure. We have recruited 300 workers from Amazon Mechanical Turk. We set the base payment to be \$0.50. Workers could earn additional bonuses depending on their performance. The average hourly rate was around \$11 USD. The experiment contains two phases as described next.

Phase 1: Learning human behavior descriptors. The goal of the first phase is to examine whether we could learn accurate human behavior descriptors from worker’s response data. In this phase, we recruited 100 workers, and each worker completed 20 rounds of product purchasing decisions. The parameters of each decision (prior, sender utility, receiver utility, and policy) was drawn uniformly at random. We split the collected data into training/test

Table 4.4: Test accuracy of different human behavior descriptors in human-subject experiments of Phase 1.

Model	Bayesian rational	TH-Model	Neural network
Testing Accuracy	0.562	0.735	0.770

sets, with 80% of the data for training, and 20% for testing. We trained and examined the performance of three different human behavior descriptors.

- Bayesian rational: This descriptor makes the standard assumption that humans are Bayesian rational. There is no training needed for this descriptor.
- TH-Model: We fit the parameters of the TH model, as described in Section 4.2.2, from data to minimize the least squares error.
- Neural network: We use a 3 fully connected-layer neural network to fit the data in the training set. We further split the training dataset and use 25% of the data as the validation set to implement early-stopping during training.

We then examine how accurately each descriptor predicts human behavior in the test data. The test accuracy is reported in Table 4.4. As we can see from the results, the data-driven neural network model leads to the best prediction accuracy, and both TH-Model and the data-driven descriptor significantly outperform the Bayesian rational assumption, reaffirming the need to relax this common assumption.

Phase 2: Evaluating HAIDNet. In the second phase, we recruited 200 workers to examine the performance of different information policies. In particular, we examine the following four information policies:

- Random: This information policy is drawn from a Dirichlet distribution.
- BR-policy: The optimal policy when the receiver is a Bayesian rational receiver.
- TH-policy: The optimal policy when the receiver behavior follows the TH-Model, as in Section 4.2.2.
- HAIDNet: The policy by HAIDNet when we use the neural network learned from the first phase as the human model.

When each worker arrives, they are randomly assigned to one of these four policy treatments. They are then presented with 20 rounds of purchase decisions (the parameters of

each round are randomly drawn from distributions fixed across all treatments) coupled with the associated information policy in the treatment. We then measure the average sender utility in each treatment. The results, as shown in Figure 4.4, demonstrate that HAIDNet achieves the best performance. The results showcase the effectiveness of HAIDNet coupled with data-driven human behavior descriptors.

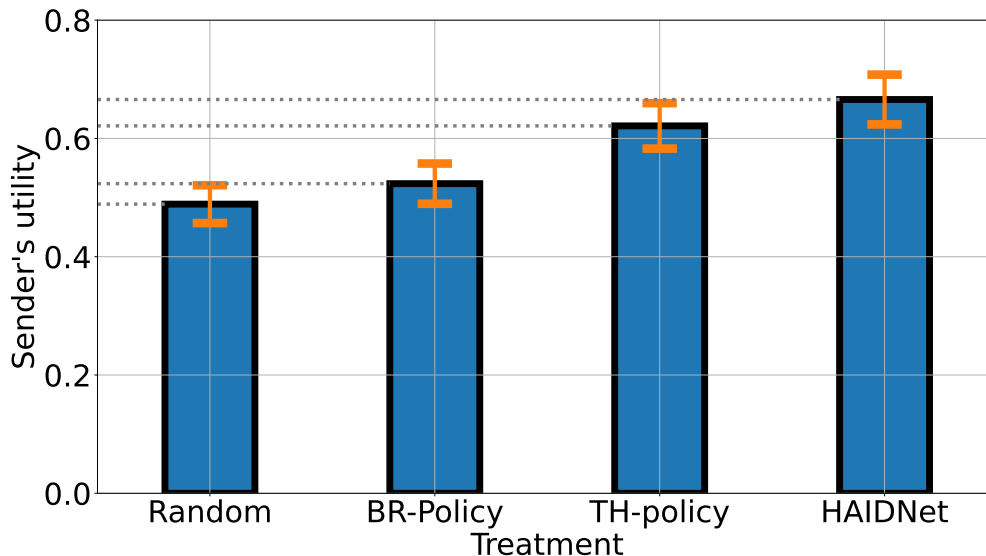


Figure 4.4: Average sender utility of different policies in human-subject experiments of Phase 2. The differences between BR-policy and TH-policy and between BR-policy and HAIDNet are statistically significant ($p < 0.01$).

In addition to examining the sender’s utility, we also measure the receiver’s utility in each treatment. We observe that, the policy of HAIDNet leads to an average receiver utility of 0.532, which is the lowest of all four treatments. This creates the concern that when we incorporate the knowledge of receiver behavior to optimize the sender’s utility in information design, we are potentially exploiting the knowledge of receiver behavior and hurting the receiver. We offer more discussion on this concern in the next section.

4.4 Discussions

We study the problem behavioral information design and encode human behavior into the design process. We propose HAIDNet, a neural-network-based optimization framework for information design that can adjust to multiple forms of human behavior. Through extensive

simulations and human-subject experiments, we demonstrate the effectiveness of HAIDNet in response to different human behavior descriptors. Below we discuss the generalization / limitations and highlight the potential social impacts.

Generalization and limitations. While this work has focused on integrating human behavior in automated information design, we believe the methodologies are generalizable to design mechanisms for general human-in-the-loop systems, explicitly encoding realistic human behavior and/or human responses to the system when designing the system. Moreover, our current investigations have adopted the most standard deep learning setup (e.g., full-connected neural networks coupled with stochastic gradient descent). It would be interesting to examine whether the performance could be further improved with carefully crafted network architecture and optimization procedure.

We would like to note the potential limitations of this approach. The optimization procedure, based on applying stochastic gradient descent on neural networks, does not guarantee to lead to globally optimal solutions in general. Therefore, it is important and interesting to explore whether and when this approach might be faced with the local optimum issue to understand the limitations and power of this approach. Moreover, compared with analytical solutions that are guaranteed to be optimal for all problem instances if the receiver behavior follows the assumption, HAIDNet is a data-driven approach that optimizes the *expected* utility, which requires training data to be representative to ensure generalizability. While our results suggest that HAIDNet recovers the near-optimal policy (e.g., the results in Figure 4.2), examining the impacts of different training data distributions and whether the results are robust to distributional shifts are potential important future research directions.

Another limitation pertains to the scalability of our proposed approach. While our method exhibits better scalability than exact solvers that utilize linear programming (more detailed discussion is included in Appendix B.2), our current results primarily focus on discrete action/state spaces. As the number of states and actions expands, so does the input size for HAIDNet. It could require much more training iterations to reach convergence. Furthermore, in scenarios with continuous action/state spaces, our approach is not immediately applicable. While discretization might be employed to address the setting with continuous spaces, such an approach requires additional smoothness assumptions to ensure small discretization errors. Overall, understanding and improving the scalability of HAIDNet is an important next step for increasing its practical applicability.

Potential negative social impacts. Finally, we highlight the potential negative social implications of the usage of information design frameworks. In information design, the sender often represents the party in power (e.g., the government, social networking platforms), while the receiver is in a less advantageous position (e.g., the general public, users) due to the asymmetry of information access. While it is possible to use information design for social good, guiding the receiver towards actions that are beneficial for himself or the public, the vast majority of information design literature — including our work — focuses on optimizing the sender’s utility. When the interests of the sender and receiver are not aligned, optimizing the sender’s utility could result in a negative impact on the receivers, who are often the general public. In other words, with an ill-specified objective in information design, the sender could exploit the information advantage and create significant negative social impacts. This concern is further amplified when we obtain more accurate knowledge about the receiver. It is therefore important to consider the impacts and potential regulations on information design.

In light of the concerns raised, to initiate the discussion, we discuss two potential risk mitigation methods. Firstly, on the technical front, we could employ differential privacy techniques [52, 51] to control the amount of private human behavior being incorporated into receiver models. Differential privacy provides a means to balance privacy with utility, typically by introducing controlled noise into the data. This mechanism might be helpful in mitigating the exploitation of marginalized groups, an issue that might be exhibited in our approach. Secondly, from a policy perspective, once we develop a comprehensive understanding of the capabilities of information design with data-driven human models, we, as a society, could and should weigh the utility gains from this method against potential harm. This discussion could then pave the way for the development of regulations and policies for deploying information design. For instance, we might impose constraints ensuring that the deployed information policy does not significantly reduce receiver utility, especially when compared to policies designed assuming standard models such as Bayesian rationality.

Chapter 5

Offering Predictions and Suggestions

In the previous Chapter 3 and Chapter 4, we assumed that the utility functions of both humans and AI systems were given when designing AI systems. However, in real-world scenarios, especially in ethically-sensitive domains such as healthcare and social services, these utility functions are difficult to define. To better understand the influence of AI-generated information on human decision-making, we focus on a medical resource allocation problem. In this context, we first learn human preferences through collected data and then utilize AI-generated information to assist humans. This Chapter is based on joint work with Saumik Narayanan, Wei Tang, Chien-Ju Ho and Ming Yin [128, 127]. I contributed to this work by designing a portion of the experiments, performing part of the data analysis, and engaging in the discussion of the results.

More specifically, we explore how predictive information or recommendations will influence human ethical decision-making. As the capability of artificial intelligence increases, AI systems are increasingly involved in decision making in high stakes domains, such as medical decision making [30, 182, 184, 117, 131], loan applications [23, 68, 98], or legal systems [6, 19]. Meanwhile, the growing prevalence of AI in decision making has raised ethical concerns, as the decisions made by these systems might be biased or might not align with human values [96, 16, 133, 6]. To address these concerns, we would ideally want to have a set of rules specifying what it means for a decision to be *ethical* such that AI researchers and practitioners can incorporate these rules when designing and deploying AI in practice. However, in ethically-sensitive domains, there are often no clear-cut right and wrong decisions. Instead, we are often forced to choose the “lesser of two evils”, prioritizing and trading off different ethical values and principles. Moreover, different stakeholders may have different preferences on the priority of ethical principles. Finding a trade-off between ethical principles that everyone agrees on for a given task may be challenging or even impossible.

To explore the above challenges and align the design of AI with human values, one natural approach is to elicit human preferences on ethical principles from relevant populations and incorporate the elicited information in the design of AI systems [105, 8, 132, 65]. In this line of work, during preference elicitation, human participants are presented information on hypothetical scenarios involving moral dilemmas and asked to express their preferences in the scenario. For example, [8] considers the moral dilemmas faced by autonomous vehicles; participants were given hypothetical scenarios in which a vehicle is bound to crash, and were asked to express their preference on sparing the lives of one group of people over another. By varying the demographics and attributes of the two groups, researchers can infer which ethical values (e.g., sparing lives, sparing youth, etc) the population prioritizes. To focus on the trade-offs in the moral dilemmas, the information presented to participants in most prior work has been *verifiable*, meaning that the information only describes the past or present, and there is no uncertainty associated with the presented information. In the meantime, as *predictive* information, which concerns predictions made about the future, is increasingly integrated in ethical decision making (e.g., judges might utilize predictive risk scores in making bail decisions), it is important to understand the influence predictive information has on human ethical preferences.

In this chapter, we aim to understand how the elicitation of human ethical preferences are impacted by the information shown to humans. We provide a contrast between verifiable information (e.g., patient demographics or blood test results) and predictive information (e.g., the probability of organ transplant success). As predictive information, from either AI or human experts, is increasingly integrated in ethical decision making, we investigate how the *presence* and the *source* of the predictive information affect human ethical preferences. We further advance our understanding of incorporating AI recommendations in ethical decision-making. We investigate how value similarity between humans and AI affects the human decision makers' reliance on AI recommendations in the context of AI-assisted ethical decision making.

To answer these questions, we conducted randomized online experiments on Amazon Mechanical Turk. Using the domain of kidney transplants as a case study, we presented scenarios where two candidates needed a kidney transplant but only one was available, and asked MTurk workers to express their preference on which candidate should receive the kidney first.

We designed three sets of experiments. In the first experiment, we investigated how ethical preferences varied between workers who saw only verifiable information and workers who saw both verifiable and predictive information. We find that even when predictions are equal between candidates, the presence of predictions change human ethical preferences. We also find that both the direction and magnitude of differences in predictive information is relevant and important for understanding how human ethical preferences change. In the second experiment, we analyzed how human ethical preferences change based on the source of the predictive information. We find that humans rely more on predictions from AI than predictions from a human doctor, possibly indicating that humans trust AI predictions more than human predictions. Moreover, humans seem to discount the importance of other verifiable information more when an AI prediction is presented, implying that humans are more likely to treat AI predictions as a summary of other verifiable information. In the third experiment, we measure the ethical preference of the participants, and design AI systems that are similar or dissimilar from the participant’s own ethical preference. We compare participants’ decision alignment with the AI recommendation across the two treatments to understand how human-AI value similarity impacts human reliance on AI. We find that recommendations provided by a dissimilar AI has a larger effect on human decisions than recommendations from a similar AI. However, this result is generally due to the high levels of agreement between the similar AI and user, creating less opportunities to “change their mind”. If we limit our analysis to the subset of scenarios where humans and AI disagree, humans are more likely to change their decision when provided with recommendations from a similar AI than recommendations from a dissimilar AI.

5.1 Related Work

The work in this chapter joins recent research in incorporating human preferences into AI systems and investigating value similarity between humans and AI systems, and we use kidney transplant allocation as a case study for human ethical decision-making.

Societal resource allocation. In this work, we study the problem of resource allocation, especially kidney transplant allocation. There has been a rich body of literature on developing algorithms for societal resource allocation [107, 43] and the associated ethics considerations [66, 54, 53, 136]. Taking from this literature, there have been a few algorithmic

experiments understanding human values for kidney allocation. [64] create a methodology for estimating human values for kidney allocation, and proposed kidney exchange algorithmic improvements which better take into account human values. For example, the United Network for Organ Sharing published a report detailing changes they made to their kidney algorithm in the last year, and showed that outcomes are now more equitable for racial minorities and other vulnerable groups [150].

Eliciting and incorporating human ethical preferences. There has been a line of research in aligning the design of AI systems with human values. One natural way to approach this alignment is to elicit real human ethical preferences in scenarios where multiple ethical principles conflict, to determine the relative weights of the principles and to understand in which scenarios, one principle might be favored over another. Correspondingly, there has been a line of work researching the elicitation of human ethical preferences [8, 65, 154, 27]. Among these works, [8] study human preferences on autonomous driving when faced with an adaptation of the trolley problem, and learned how these ethical preferences vary across worldwide cultures. [166] study human preferences in moderation of Wikipedia quality prediction. [65] study human preferences in the allocation of kidneys for transplants. Our work differs from this line of work in that we focus on discussing the impact of predictive information to human ethical preferences while existing work have mostly utilized verifiable information only. Another related work by [27] also analyze the elicitation of ethical preferences in the kidney domain. However, they analyze how assessments of human ethical preferences impact their ethical decision making, and don't focus on the impact of predictive information to human ethical preferences. As a closely related line of research, if we consider different fairness measures as different ethical principles, our work is also related to the research in understanding human perceptions of different fairness measures [76, 168, 187, 183], especially because it's usually impossible to satisfy all fairness measures simultaneously [21, 35, 99, 32].

Some recent research focus on utilizing participatory design to govern the design and implementation of AI systems [105, 132, 202, 166]. These works look at the next steps after we have elicited these ethical preferences, namely how to integrate these preferences into the deployment of the AI systems. For example, [202] look at methods of presenting these preferences to stakeholders, so that they better understand the trade offs that they must make. [132] construct a system where multiple models of ethical preferences vote on which principles should be used for a given scenario, based on pre-elicited human preferences, and

[105] explore how such a participatory framework could leverage multiple stakeholders during the decision-making process.

Human reliance on AI systems. In studies of humans' reliance on AI advice, there have been mixed results on whether humans rely more on human advice or AI advice. Many papers have shown evidence of algorithmic aversion, which is the notion that humans tend to relatively distrust AI advice, and prefer to receive advice from other humans [143, 40, 113]. This aversion extends to second and third parties, who may prefer decision-makers to use no advice, rather than AI advice [194, 158]. On the other hand, despite the evidence that decision-makers tend to subjectively prefer human advice over AI advice, [112] find that human-decision makers tend to rely more on AI advice in practice. This finding has been validated not only in objective domains, but ethical decision-making domains where there are no correct answers [128, 177]. One potential explanation is that humans perceive AI to be more rational and unbiased [41]. Human decision-makers may also want to shift the cognitive burden of ethical decision making off of them [86], as society tends to hold humans to higher standards of being unbiased than AI [20]. One aspect which affects human reliance on AI is trust, or more generally, the level of confidence that humans have in AI outputs. [15] investigate the mental models that humans have in AI behavior, and find that when model outputs are more understandable, humans are better able to incorporate these outputs into their own decision-making strategies, leading to better team performance. [200] investigate the relationship between model accuracy and trust, and reveal that humans tend to both trust and rely on advice with a higher stated accuracy more than advice with a lower stated accuracy. [156] find that when humans are exposed to AI advice and later show that the prior advice is incorrect, their trust in the AI actually increases. [213] look at methods for calibrating human trust in AI, and show that confidence scores improve trust calibration, though this doesn't necessarily improve overall decision making performance.

Value similarity between human decision-makers and AI systems. We also study the effects of value similarity to human reliance in AI-assisted ethical decision making. There is a rich body of sociological work understanding the effects of value similarity on humans. For example, [163] find that improving reliability is insufficient for restoring trust in interpersonal relationships or inter-organizational mechanisms, and a better method for improving trust is to show value similarity. [161] analyze the effects of value similarity in risk management, and show that increased value similarity leads to increased trust and is a significant

predictive factor in the outcome of risk-benefit analysis for new technology. One of the closest work to ours is by [75]. They focus on objective (non-ethical) domains and measure AI similarity by comparing model output with human decisions. Similar to our observations, they find that advice from similar AIs is more likely to change the mind of a human decision maker, but dissimilar AIs have more opportunities to change minds, giving them a bigger overall impact. [125] and [201] both investigate the effects of value similarity on AI trust in various ethical decision-making domains, and find that AI assistants with a higher value similarity lead to higher levels of trust in the AI assistant. However, the latter two papers only look at subjective measures of trust in these ethical decision-making domains, without empirically validating changes in user reliance. We have already seen paradoxical results when looking at reliance on human and AI advice, where decision-makers prefer and trust human advice more, but rely on AI advice more. As such, we aim to fill this research gap in AI-assisted ethical decision-making, by showing that value similarity in AI recommendations leads to both increased reliance and increased trust.

5.2 Problem Formulations and Research Questions

In this work, we use the domain of kidney transplants as a case study. There has been extensive literature on the ethical principles in allocating scarce medical interventions [137, 55, 53, 67]. In particular, our task design is based on the work by [137], who list the following four categories of ethical principles for allocating scarce medical resources.

- Promoting and rewarding social usefulness: This principle could be implemented through prioritizing *instrumental value*, e.g., giving medical workers higher priority in receiving vaccines during a pandemic, or *reciprocity*, e.g., giving prior organ donors higher priority to receive a transplant of their own.
- Treating people equally: In this principle, everyone should have equal chance of receiving medical interventions. It can often be implemented using *lottery* or *first-come-first-serve* approaches.
- Favoring the worst-off: This principle could be implemented through deploying the strategy of *sickest first*, prioritizing those who have a more severe disease condition or *youngest first*, prioritizing those who have not lived as many years yet.

- Maximizing total benefits: This principle aims to maximize some definition of utility, e.g., maximizing the number of saved lives or maximizing the increase life-years after intervention.

These categories of ethical principles are widely used, both in academic contexts [53, 190, 102, 137], and in action for real-world medical organizations [151, 139].

We explore three research questions about the effect of AI systems in ethical decision-making:

- **Research Question 1:** How does the presence of predictive information affect human ethical preferences?
- **Research Question 2:** How does the source of the predictive information (e.g., predictions by human experts or predictions by AI systems) affect human ethical preferences?
- **Research Question 3:** How does value similarity affect human reliance on AI recommendations?

To study these problems, we recruit workers to make decisions in a set of kidney transplant scenarios. In each scenario, workers are presented two patient candidates who both need a kidney transplant, but only one kidney is available. Given information about each of these candidates, workers are asked to express their preference on which candidate should receive the kidney first. Based on the ethical principles which govern the allocation of scarce medical resources [137], we choose four factors to display to workers. The first three factors concern the present condition and attributes of the candidates, which we denote as *verifiable information*. The fourth factor concerns a future prediction made about the candidates, which we denote as the *predictive information*. Specifically, these factors (along with the corresponding ethical principle) are:

- **Kidney Donor Status** (Promoting social usefulness): Whether the candidate has donated a kidney of their own in their past. This is a binary feature, with possible values of {Not prior donor, Prior Donor}.
- **Wait Time** (Treating people equally): How long the candidate has been waiting to receive a kidney transplant. This feature has possible values of {Less than 1 year, 1 year, 2 years, 3 years, 4 years, 5 years}.

- **Kidney Disease Stage** (Favoring the worst-off): How severe the candidate’s kidney disease is. This is a binary feature, with possible values of {Stage 4 (Severe kidney damage), Stage 5 (Kidney failure or near-failure)}.
- **Post-Transplant Survival Chance** (Maximizing total benefits): The predictive probability that the candidate will remain alive after 5 years post-transplant. This feature has possible values between 72% and 98%.

Based on the established ethical principle framework [137], there is a preference ordering on each factor when all other factors are equal. For example, if two candidates share the same values for kidney donor status, kidney disease stage, and post-transplant survival chance, the patient with longer wait time is preferred according to the ethical principle. In our experiments, we present different scenarios to online workers to understand how humans make trade-offs on these four factors, mapping to the four corresponding ethical principles. One example of the experiment task is shown in Figure 5.1, and detailed experiment interfaces are discussed in Appendix C.3.

Question 10 of 29

Which candidate should receive the kidney transplant first?

	Patient A	Patient B
Kidney Donor Status	Prior Kidney Donor	Not Prior Donor
Wait Time	1 year	3 years
Kidney Disease Stage	Stage 5 (Kidney failure or near-failure)	Stage 5 (Kidney failure or near-failure)
Post-Transplant Survival Chance	74% chance of survival after 5 years	76% chance of survival after 5 years
	Select: Patient A	Select: Patient B

Please make your selections.
Click the buttons or use the ←/→ keys.

Figure 5.1: Human experiment interface of ethical decision-making in kidney allocation with AI generated predictive information. The information of post-transplant survival chance is generated by AI systems.

5.3 Experiments

We conduct three experiments to answer our research questions, and additional results are available in Appendix [B.3](#).

5.3.1 Experiment 1: the Effect of Predictive Information

To understand the effect of predictive information on human ethical preferences, we conducted a randomized behavioral experiment with two treatments.

- **Treatment 1 (Verifiable Only):** This treatment group is shown the three factors of verifiable information. This represents the human priors on human ethical preferences, and gives us a baseline to measure the effects of the predictive factors against.
- **Treatment 2 (Verifiable and Predictive):** The treatment group is shown both the three verifiable factors, and one factor based on predictive information. We did not present the source, explanation, or any other information about this predictive factor.

Each recruited worker is asked to express their ethical preference in 29 scenarios. In each scenario, workers are presented with two candidate profiles and are asked to provide their preference on which candidate should receive the kidney transplant first. We show an example of what a worker in the second treatment (verifiable and predictive) see in [Figure 5.1](#). Workers in the first treatment (verifiable only) see the same design, except they are not shown the predictive information of post-transplant survival chance in the last row.

Scenario selection. In the first treatment (verifiable only), workers are only presented verifiable information about the candidates. Each of the three verifiable factors are ordinal, and we have two candidates presented in each scenario, which we label as A and B. This gives us three possible orderings for each factor: candidate A is preferred over candidate B, candidate B is preferred over candidate A, and both candidates are equally preferred. Because we have three factors and three orderings, we get 27 total scenarios of factor orderings to assign. We discard the one scenario where both candidates share the same values for all factors and are left with 26 scenarios. Each worker in the first treatment group will view each of these 26 combinations once. Each combination is realized with randomly generated values. If we want donor status to be equal, we may display both patients as "Prior Kidney

Donor”, or both ”Not Prior Donor”. If we want the wait time of A to be higher than B, we may show 2 years and 1 year, or 5 years and 3 years, or any other pair of values as long as the difference is no more than two years. After the worker views the first 26 scenarios, we randomly choose three of the scenarios shown and show these scenarios to the worker again, with the exact same realization of the factor values. We do this as a consistency check, so we can determine the quality of a particular worker’s data by how consistent their preferences are over these three repetitions of scenarios. To minimize the potential presentation bias caused by the ordering of the scenarios, we randomize the first 26 scenarios. To minimize the potential bias caused by the ordering of the candidates, we randomize the order of the candidates independently for each scenario.

In the second treatment (verifiable and predictive), workers are presented both verifiable and predictive information about the candidates. Note that the additional predictive factor is also ordinal, with three directions. Each worker is also presented 29 scenarios. To generate the combinations for the second treatment group, we take the same 26 combinations as in the first treatment, but when we present this to workers, we randomly select a direction for the predictive information (whether the predicted survival chance of one candidate is larger than, equal to, or smaller than the other), and show this to workers. As with the wait time feature, we randomly select a pair of values for each scenario, where values can be between 72% and 98%, and constrain the difference to be no more than 6%. We then again add three repeated scenarios randomly drawn from the first 26 scenarios for consistency check. We then apply the randomization procedure for the ordering of the first 26 scenarios and the presentation order of the two candidates in each scenario.

To answer the first research question, we compare workers’ preferences in the first treatment with workers’ preferences in the second treatment on the scenarios where the two candidates have the same predicted survival chance. Given the number of scenarios in the second treatment for the above comparison is only one-third of the number of scenarios in the first treatment (as we randomly draw the ordering of predictive information from the three possible orderings), during random treatment assignment, we assign three times more workers in the second treatment compared with the number of workers assigned to the first treatment. We further split the workers’ preference data collected from the second treatment into three groups, based on the direction of predictive information, and analyze how this direction affects their ethical preferences.

Experiment procedure. For this experiment, we recruit participants by posting a HIT on Amazon Mechanical Turk. The HIT is only open to U.S. workers. In the preview page of the HIT, workers first view an instruction page, a sample scenario, and the consent form. Workers need to agree to the consent form to accept the HIT and participate in the experiments. After accepting the HIT, workers are randomly assigned to one of the treatments, with 25% chance of being assigned to the first treatment and 75% chance of being assigned to the second treatment. Workers are then shown a background page explaining the factors used for determining which candidate would receive a kidney. Workers are only presented the explanations on the factors used in their corresponding treatments. Afterwards, the workers begin to evaluate kidney transplant scenarios. While evaluating scenarios, workers are still able to reference the background information on transplants. Finally, workers are asked to complete a short demographic survey.

Performance measure. To measure workers’ ethical preferences from collected data, we use conjoint analysis to compute the average marginal component effect (AMCE) of each factor (kidney donor status, wait time, kidney disease stage, and post-transplant survival chance). More concretely, for each factor, we select all scenarios where the factor value is unequal, and aggregate the average number of times that workers select the higher value over the lower value (recall that for each factor, there is an ethically preferred direction). We calculate the percentage of workers who select the higher value and the percentage of workers who select the lower value, and denote the difference between these values as ΔP . For example, to calculate ethical preferences for the kidney donor status, we select all scenarios where one patient is a prior kidney donor and the other patient is not, and measure the difference between the preference of the former and the preference of the latter. This difference is the reported ΔP .

We recruit a total of 600 workers, with 184 workers being assigned to the first treatment, and 416 workers being assigned to the second treatment. We discard workers who are not completely consistent on the three consistency check questions and report the results for the 202 workers who are fully consistent. We have also performed the same analysis on the entire worker pool, and the results are qualitatively the same.

The effect of equal prediction on human ethical preferences. We first examine whether the addition of equal predictions between candidates have any effect on human

ethical preference compared with no predictive information. We compare the ethical preference from the first treatment (verifiable only) and the ethical preferences from the subset of samples with equal values in the predictive factor in the second treatment (verifiable and predictive).

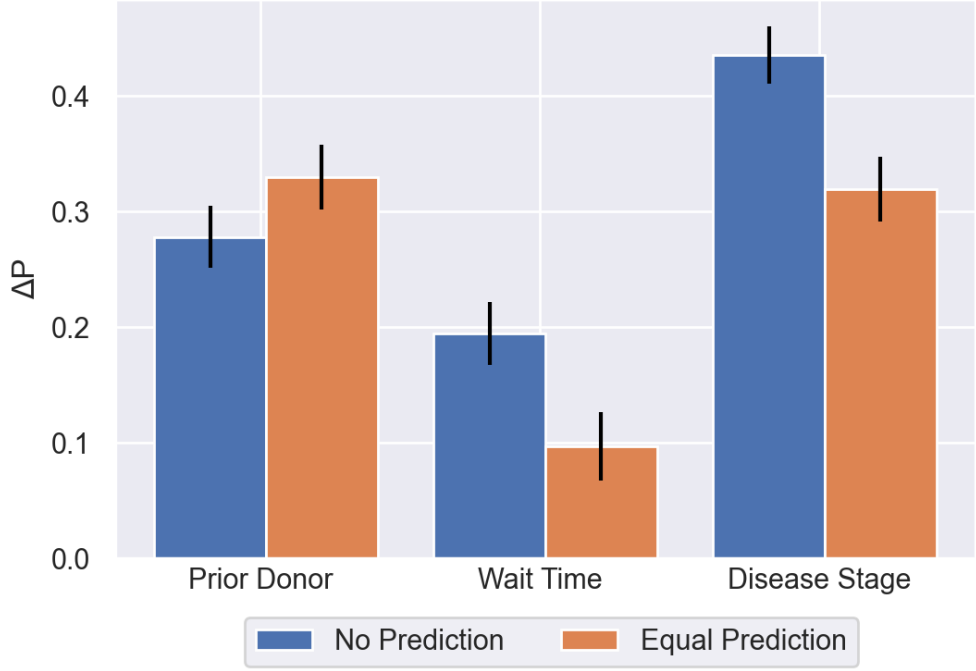


Figure 5.2: The effect of equal prediction on human decisions. We present ΔP for each verifiable factor and treatment. There is no significant difference between treatments in the Prior Donor factor ($p = .54$). There is a significant difference between treatments in the Wait Time factor ($p = .045$). There is a significant difference between treatments in the Disease Stage factor ($p = .0057$).

The results are shown in Figure 5.2. We compare ΔP (the difference between preferring the higher value in a factor and preferring the lower value in a factor) for the three factors in verifiable information between the first treatment and the second treatment where the predictive factor is equal between candidates. We also apply Bonferroni correction to our significance tests to account for multiple comparisons. The first treatment represents the baseline of human ethical preferences when no predictive information is available, and the second treatment represents situations where predictive information is shown to humans, but does not favor either candidate. We find that the presence of equal predictive information significantly decreases the ethical preference of Wait Time from 0.194 to 0.097 ($p = .045$), significantly decreases the ethical preference of Disease Stage from 0.435 to 0.319 ($p = .0057$),

and increases the ethical preference of Prior Donor from 0.278 to 0.330, though this increase is not significant ($p = .54$). These results show that human ethical preferences do change even when predictive information is presented and is equal across candidates.

Interestingly, these changes are not consistent for all ethical preferences. We speculate that the reason for this is because humans may create their own predictions about the scenario based on the verifiable information we present, but when we present an externally sourced prediction about the scenario, they no longer create their own predictions and instead use the prediction provided. For example, one possible conjecture for the explanation of the result is that workers might think wait time and disease stage is more predictive of survival outcomes than prior donor status. Therefore, workers in the first treatment without predictive information may have used these in forming their own predictions which influence their ethical preferences. But when we present the prediction, this supersedes their own prediction, and their final preference is weighted less heavily towards wait time and disease stage when predictive information is available.

The effect of aligned prediction on human ethical preferences. We next examine whether the addition of predictions strengthens human ethical preferences if the predictions are aligned with the preferences. The results are shown in Figure 5.3, in which we compare the difference in ΔP in each factor based on the three possible directions of prediction alignment from the samples in the second treatment. We also apply Bonferroni correction to our significance tests to account for multiple comparisons. For each factor, we first select all scenarios where the factor value is unequal in the second treatment. We then split the samples into three groups (Aligned, Equal, or Misaligned), depending on how the preference of the prediction aligns with the preference of the verifiable factor. We then calculate the values of ΔP , the difference between the ratio of workers choosing the higher value and the ratio of workers choosing the lower value, for each factor and each group. We find that for all factors, there is a significant ($p < .001$) difference between misaligned prediction and equal prediction, and that there is a significant ($p < .003$) difference between equal prediction and aligned prediction. These results show that human ethical preferences are strengthened when predictions are aligned with the human preferences, and weakened when predictions are oppositely aligned with the preferences.

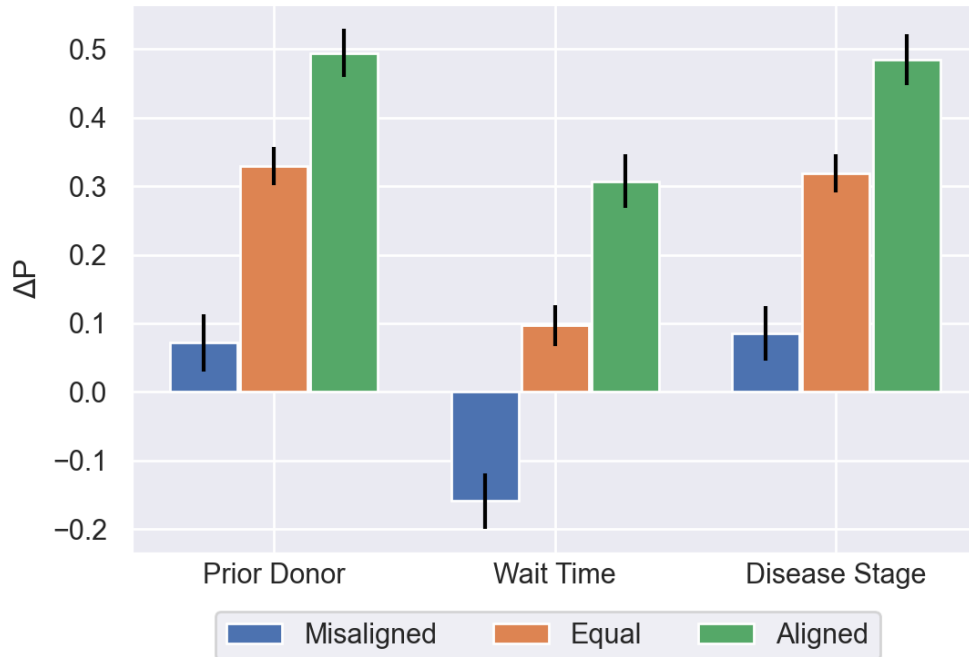


Figure 5.3: The effect of aligned prediction on human decisions. There is a significant difference between a misaligned prediction and equal prediction for all factors ($p < .001$). There is a significant difference between a equal prediction and aligned prediction for all factors ($p < .003$).

5.3.2 Experiment 2: the Effect of Prediction Sources

In the second experiment, we investigate our second research question: how the effect of predictive factors on human ethical preferences changes based on the source of the prediction. Specifically, we aim to find if there are differences if we tell workers that the prediction is generated by a human doctor or an AI system. Similar to the setup of Experiment 1, we recruit works to making decisions in different scenarios with predictive information, but at this time we also presenting the resource of the prediction. In order to examine whether there are differences if we tell the user that the prediction is generated by a human doctor or AI, we create two treatment groups with varying prediction sources:

- **Treatment 1 (Doctor):** The first treatment group is shown the three demographic factors, the predictive factor, and an explanation saying that the prediction is generated by a human doctor.

- **Treatment 2 (AI):** The second treatment group is shown the three demographic factors, the predictive factor, and an explanation saying that the prediction is generated by an AI system.

Each recruited worker is asked to express their ethical preference in 29 scenarios. The choice of the 29 scenarios is the same as the second treatment in Experiment 1, with the addition of the prediction source, which is given along with the predictive value. The first 26 scenarios reflect all combinations of factors in verifiable information and a random draw of predictive information. The last three scenarios are randomly drawn from the first 26 for checking worker consistency.

We recruit a total of 300 workers, with 156 workers being assigned to the first treatment, and 144 workers being assigned to the second treatment. We discard workers who are not completely consistent on the three consistency check questions and report the results for the 127 workers who are fully consistent. We have also conducted the same analysis on the entire worker pool, and the results are qualitatively the same.

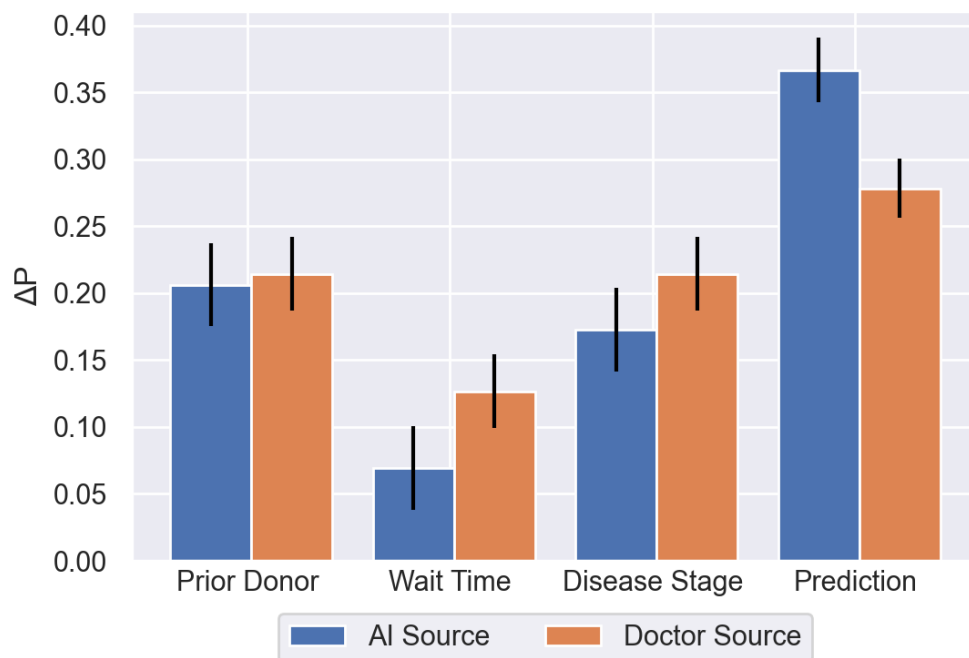


Figure 5.4: The effect of prediction source on human decisions. There is no significant difference between treatments in the Prior Donor factor, Wait Time factor, or Disease Stage factor. There is a significant difference between treatments in the Predictive factor ($p = .0316$).

In Figure 5.4, we see how ΔP changes based on the source of the prediction for each human ethical preference factor. We also apply Bonferroni scaling to our significance tests. We see that changing the prediction source from AI to Doctor significantly decreases the ethical preference of the prediction ($p = .0316$). From this result, we actually see evidence suggesting that human ethical preferences from a prediction are weakened when the prediction source is a human doctor, and strengthened when the prediction source is an AI. We see that changing the prediction source from AI to Doctor increases the preference of Prior Donor, Wait Time, and Disease Stage, though not significantly. Combining both observations, one plausible conjecture is that workers might believe that AI predictions are generated by incorporating all verifiable information. Therefore, their preferences are influenced more by AI predictions instead of doctor predictions. Moreover, when AI predictions are available, workers put a smaller weight on other factors as they might be incorporated in AI predictions already. Our results suggest that how humans process predictions might vary when the predictions are from different sources.

Exploratory analysis. In our post-scenario survey, we ask workers to report the perceived trustworthiness of the predictive information on a five-point scale, as well as demographic information on age, gender, race, education level, and political leanings. We find that workers in the doctor treatment rate perceived trustworthiness of the prediction as 1.85/5, and workers in the AI treatment rate the perceived trustworthiness of the prediction as 1.96/5. This aligns with our results which show that human preferences are more influenced by AI predictions than human predictions, and prior literature which suggests that humans trust AI more than human experts [112]. It is interesting to note that the relative values of perceived trustworthiness are so low for both, especially considering that the workers involved are layperson, and prior research shows that experts trust algorithms less than layperson do [111, 112, 106].

We find that perceived trust is negatively correlated with levels of education. Workers with a bachelor's degree report 0.37 lower perceived trust in predictions than workers without a bachelor's degree. Workers with a graduate degree report 0.36 lower perceived trust in predictions than workers with a bachelor's degree or lower. This trends hold when we split workers by treatment (Doctor vs AI). We speculate that cause for this trend is that as humans believe themselves to be more capable, they tend to rely less on advice from others [45].

In our main analysis, we analyze the difference in ΔP values between the AI prediction and Doctor prediction. For context, the total pool of workers have an average ΔP difference of 0.089. This value can be considered as a proxy of the difference between humans' reliance on AI prediction and the reliance on doctor prediction. To understand whether there exist individual differences, we break this down by demographic. We find that workers above the age of 40 have a ΔP difference of 0.027, while workers below the age of 40 have a ΔP difference of 0.122, suggesting that the majority of difference in overall workers is based on age, where younger workers' preferences are more influenced by AI predictions than doctors' predictions, compared to older workers. We find that male workers have a ΔP difference of 0.072, while female workers have a ΔP difference of 0.080, which does not suggest a strong contrast according to gender. We find that liberal workers have a ΔP difference of 0.058, while conservative workers have a ΔP difference of 0.101. Interestingly, conservative workers have higher values of ΔP than liberal workers regardless of source, with ΔP values of 0.407 and 0.276, respectively. While the presented results are not causal, the results as a whole suggest that there are individual differences in how humans incorporate AI/doctor predictions, and it would be an interesting future direction to further explore these individual differences.

5.3.3 Experiment 3: the Effect of Value Similarity

In our third experiment, we explore the effect of the value similarity between AI systems and humans. Similar to previous two experiments, workers are asked to express their ethical preference on which candidate should receive a kidney transplant first. But instead of presenting predictive information from AI systems, we directly generate AI suggestions on final decisions.

When eliciting workers' ethical preferences, these scenarios can be split into three categories. The first category includes scenarios where the two candidates differ in only one factor, and share the same values for the other two factors. For example, in one scenario, Candidate A may be a prior donor, while Candidate B is not; both candidates have been waiting for 3 years and have Stage 4 Kidney Disease. The primary objective of this category is to elicit workers' baseline preferences for each of the factors individually (in this case, *Donor Status*). The second category consists of scenarios to understand workers trade-offs between two factors. In this category, the two candidates share the same value for one factor, one

factor should prioritize the first candidate, and the remaining factor should prioritize the second candidate (according to the default preference ordering). For example, Candidate A may be a prior donor, while Candidate B is not, Candidate A may have been waiting for 2 years, while Candidate B has been waiting for 4 years, and both candidates have Stage 5 Kidney Disease. This category enables us to isolate the trade-offs between pairs of factors (in this case, *Donor Status* and *Wait Time*). The third category involves scenarios where the two candidates have different values in all three factors. One candidate is prioritized in one factor, while the other candidate is prioritized by the other two factors. For example, Candidate A may be a prior donor, while Candidate B is not, Candidate A may have been waiting for 2 years, while Candidate B has been waiting for 4 years, and Candidate A may have Stage 4 Kidney Disease, while Candidate B has Stage 5 Kidney Disease. This category enables us to represent more complex interactions between the factors.

In each of these categories, there are three unique scenarios, giving us a total of nine scenarios. For each user, we realize each scenario with random values that preserve the preference order. For instance, if the disease stage needs to be equal, we may display both patients as "Stage 4" or "Stage 5". We also limit wait time differences between candidates to be no more than 2 years.

Similar or dissimilar AI treatments. Given our goal is to investigate the influence of value similarity between humans and AI on human reliance for ethical decision-making, we use the similarity of ethical preferences to represent the value similarity. We now describe how we create AI systems with similar or dissimilar ethical preferences with a given worker.

For a worker's ethical preference, we can measure their answers on a set of given scenarios, i.e., their choices on who to receive a kidney first among several pairs of candidates, when they are not provided AI recommendations. Using their answers, we can compute their (prior) ethical preferences without seeing AI recommendations. A worker's ethical preference is represented by three values, each indicating how often workers' answers align with the default ethical ordering of each factor. This alignment is measured separately for each factor, and indicates how often the worker chooses the preferred factor value (e.g. "Prior Donor" over "Not Prior Donor" for the "Donor Status" factor), across all scenarios. For example, if the worker selected Patient A, then their answer aligns with the preferred factor for the "Wait Time" and "Disease Stage" factors, but not the "Donor Status" factor. We would then average the number of times the worker aligns with each preferred factor across all scenarios to generate

the alignment values for each factor. Using these values, we use the $A > B > C$ notation to denote a worker's value ordering in their ethical preferences over factors A, B, and C. For example, if a worker aligns with the "Donor Status" factor in 30% of scenarios, with the "Wait Time" factor 80% of the time, and the "Disease Stage" factor in 50% of scenarios, then their prior ethical preference ordering would be "Wait Time" > "Disease Stage" > "Donor Status".

Based on a worker's value ordering in the prior ethical preference, we can design a similar AI and a dissimilar AI that share similar and dissimilar ethical preferences with the worker. In particular, if a worker's value ordering is $A > B > C$, the ethical preferences for the similar/dissimilar AI for that worker are specified below:

- Similar AI: The ethical preference order for a similar AI is chosen uniformly at random to be either $A > B > C$ or $A > C > B$, i.e., the top factor of the similar AI is the same as the top factor of the worker.
- Dissimilar AI: The ethical preference order for a dissimilar AI is chosen uniformly at random to be either $C > A > B$ or $C > B > A$, i.e., the top factor of the dissimilar AI is the same as the bottom factor of the worker.

Besides value similarity, we also instruct AI systems to behave in deterministic manner or random manner. The deterministic AI will deterministically follow its ethical preference ordering. If the Deterministic AI's top ethical preference has different values for the two candidates, then the AI will pick the candidate whose factor value aligns with its preference. If the values are tied, then the deterministic AI will move to the second preference, and then the third if necessary. However, the random AI chooses the recommendation entirely randomly, without any regard for the candidate attributes.

To understand the effect of AI similarity on the usage of AI recommendations in ethical decision making, we conducted a two-stage, two-treatment randomized behavioral experiment. In our experiment, each recruited worker begins with the first stage, where they are asked to express their ethical preferences in 9 scenarios, generated using the approach described above. After eliciting workers' prior ethical preferences, we then randomly assign workers to two treatments:

- **Treatment 1 (Similar AI):** In the second stage, each worker in this treatment group are shown recommendations from AI with similar ethical preferences to their own ethical preferences.
- **Treatment 2 (Dissimilar AI):** In the second stage, each worker in this treatment group are shown recommendations from AI with dissimilar ethical preferences to their own ethical preferences.

After the first stage, workers are presented with a summary of their own ethical preferences and the ethical preference of the AI that will make recommendations during their decision-making during the second stage. Workers are also asked three survey questions: how confident they are in their own answers, whether they think our estimation of their preferences is accurate, and how much trust they would have in an AI which behaves according to the displayed preferences. Each of these is graded on a 5-point Likert scale.

In the second stage, workers are presented with 18 additional scenarios where they make their decisions with the assistance of the provided AI. The scenarios are generated the same way as in the first stage, but the number of scenarios are doubled and the realizations of the factor values might not be the same. In both treatments, workers will encounter a deterministic AI in 9 scenarios, and a random AI in the other 9 scenarios. These are shuffled so workers don't know whether recommendations are deterministic or random. Because the Random AI could still pick the patient according to its original value preference ordering by chance, the combined AI (Deterministic+Random) follows its stated value preference ordering stochastically, about 75% of the time.

Once the worker finishes the second stage of the experiment, they will fill out an additional survey where we ask workers for a general demographic description, and two more questions about their experience: which dimension (Prior Donor, Wait Time, Disease Stage) most impact their decision making without the AI, and how much do they think they rely on the AI when making decisions in the second stage. We recruit a total of 300 workers, with 160 workers being assigned to the first treatment, and 140 workers being assigned to the second treatment.

We measure human reliance in two different ways. First, we express reliance as the overall change in alignment between the human and AI between the first and second stages. Then, we express reliance as the change in decision-making behavior, computed only on the subset

of scenarios where the human and AI differ in the first stage. We present results for both of these metrics in Figure 5.5. We report the statistical significance values using a t-test and the effect sizes using Cohen’s d . Error bars in plots represent standard errors.

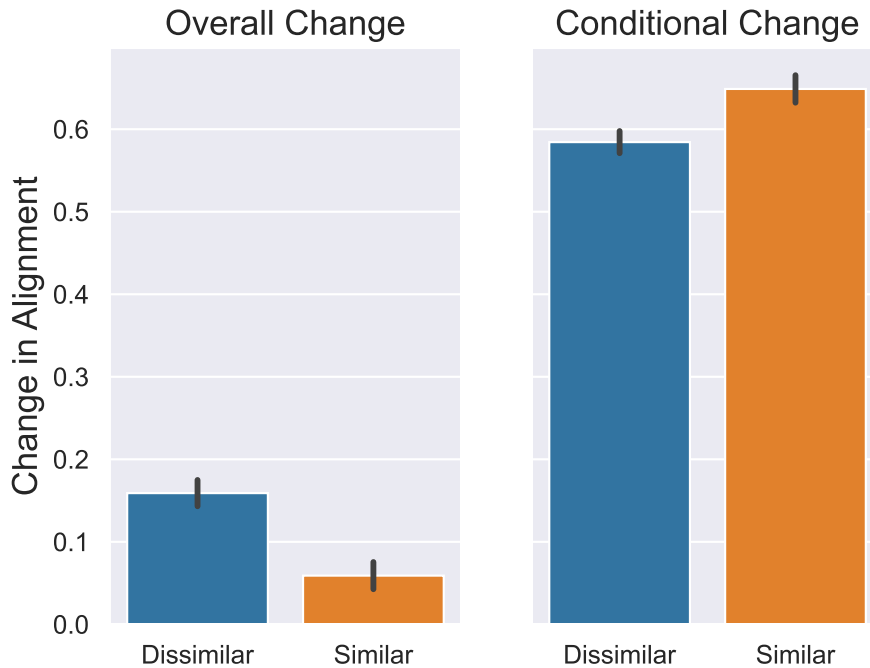


Figure 5.5: The effect of value similarity on alignment change between Stage 1 and 2. In the left figure, we find across all scenarios, the dissimilar AI has a significantly larger change in alignment ($p < .001$). In the right figure, we find that in scenarios where the human and AI disagree, the similar AI has a significantly larger change in alignment ($p = 0.003$).

Overall change in alignment. In order to measure the overall change in alignment, we compare the rate at which users match with the (unseen) AI in the first stage with the matching rate in the second stage. We find that adding a recommendation from a similar AI significantly increases alignment by 5.9% ($t(1286) = 3.58, p < .001, d = 0.10$), while adding a recommendation from a dissimilar AI significantly increases alignment by 15.9% ($t(1439) = 9.98, p < .001, d = 0.26$). The difference between the two increases is also significant with $t(2705) = 4.35, p < .001, d = 0.17$. Overall, we find that dissimilar AIs have a bigger overall impact on overall alignment. While this result may seem unintuitive, it can be explained by the fact that users tend to agree more with a similar AI than a dissimilar AI, so there is less room to increase agreement for a similar AI in the second stage.

Conditional change in alignment. As a perhaps more useful measure of reliance, we can choose to consider only scenarios where the AI gives recommendations which go against the decision that the user made in the first stage. This comparison is possible because our experiment design guarantees that each of the nine possible scenarios appear once in the first stage, and twice in the second stage. We find that when the AI gives a recommendation which goes against the user’s Stage 1 decision, alignment with a similar AI increases by 64.9%¹², while alignment with a dissimilar AI increases by 58.4%. This difference is significant with $t(1302) = -3.00, p = 0.003, d = 0.17$. Overall, we find that similar AIs have a bigger impact on human alignment when the AI goes against human prior preferences. Additional analysis is available in Appendix B.3.

5.4 Discussions

In this section, we discuss the limitations, implications, and future work.

Limitations and generalizability. Our study has a few limitations. First, our work has used the domain of kidney transplants as a case study to investigate how predictive information affects human ethical preferences. We believe this domain is representative of the family of problem domains involving allocating scarce medical interventions, e.g., organ transplants, vaccine distributions, or ventilator allocation. Relaxing the application beyond medical domains, our problem domain is in the family of domains involving allocation of scarce societal resources, such as allocating homelessness resources to people in need. We conjecture that the results of our study are very likely to generalize to the domains of medical resource allocation and are also likely to generalize to scarce societal resource allocation. However, it is also possible that our results will not directly generalize to these domains due to the uniqueness of the domain of kidney transplantation. Therefore, more future studies should be conducted to examine the generalizability of our results in other domains thoroughly.

¹²Because we are only examining scenarios where the human originally disagreed with the AI, these increases can be interpreted as total alignment in the second phase. E.g., in this subset of scenarios, workers choose to follow similar AI recommendations 0% of the time in the first stage, and 64.9% of the time in the second phase.

Implications of our results. Despite the limitations, our findings suggest a few important implications. First, our results suggest that the inclusion of predictive information impacts human ethical preferences in a nontrivial manner. Humans might consider what other factors might have already been incorporated in generating the predictive information and adjust their ethical preferences accordingly. We do not have a definite answer on how humans process predictive information. However, as predictive information is becoming increasingly involved in ethical decision making, it is important to understand how humans incorporate predictive information in forming their ethical preferences. Moreover, as shown in our exploratory analysis in Section 5.3, there exist individual differences in how people process predictive information. It is therefore important to take this into account when utilizing the elicited information to inform the design of AI systems.

Another important implication is on the robustness of elicited ethical preferences. Our results demonstrate that human ethical preferences could change significantly depending on how information is presented to them (e.g., highlighting the source of predictive information). This suggests that the elicited human ethical preferences might not be entirely robust and might be subject to information manipulation. While the growing literature on participatory design [105, 132, 202] have attempted to involve stakeholders in shaping the design of AI systems, our results suggest that, using the techniques from the literature on information design [94, 172], the advantageous party (e.g., the party that performs the elicitation) might strategically choose the information presentation to lead populations to express preferences that align with their objective. It is therefore important to understand under what conditions and to what extent we might rely on these elicited human preferences to guide the design with the goal of aligning AI with human values.

Future work. Our work has presented interesting findings on the effect of predictive information and value similarity to human ethical decision-making. However, there are still a lot of open questions that deserve future study. For example, how do human ethical preferences change when the presented predictive information becomes more accurate? If we explain how the predictive information is generated, does it impact how humans incorporate the information into their ethical preferences? Again, as predictive information becomes more ubiquitous, it is important to have a better understanding on how the presence and presentation of the predictive information impact humans. Moreover, as brought up by the above discussion on the limitations and implications, more studies on different problem domains and the populations surveyed would help us understand the generalizability of the results.

It is also important to study how to leverage this elicited information to inform the design of AI systems and whether the elicited information is robust against potential manipulations.

Chapter 6

Incorporating Human Beliefs about AI Behaviors

Previous chapters discussed how to update the decision-making environment and design the information humans will collect to influence their decision-making. In these scenarios, humans are the decision-makers, and the reward of AI systems is based on human actions. However, AI systems can also be decision-makers as humans are, with reward functions dependent on the joint actions of both AI and humans. This introduces the domain of human-AI collaboration, where a new challenge arises: humans may adjust their behavior based on the actions of AI systems. In this chapter, we explore potential methods to model human dynamic behavior and train AI systems to effectively cooperate with humans. This Chapter is based on joint work with Robert Kasumba, Chien-Ju Ho, and William Yeoh [204] (under review). I contributed to this work by proposing the incorporation of human beliefs about AI teammates, modeling human behavior and beliefs, and designing and conducting experiments to evaluate the proposed methods.

The potential for human-AI collaboration is immense and spans various domains. In health-care, AI systems can identify diagnoses that might be overlooked by human professionals [29, 126]. In industrial manufacturing, robots work alongside human workers to enhance efficiency and safety [17, 160]. In workflow productivity, virtual assistants can generate drafts for humans to refine and finalize [193]. However, despite significant improvements in AI performance over the past decade, designing AI agents to optimize the overall performance of human-AI collaboration remains a challenge.

In particular, optimizing the AI system in isolation is not sufficient to enhance the performance of human-AI collaboration. Both the human and the AI agent need to *coordinate* by inferring the goals and intentions of their counterpart and taking complementary actions. For example, in AI-assisted decision-making, researchers have demonstrated that instead of

optimizing AI by itself, training AI to focus on improving in areas where humans typically struggle can significantly enhance the performance of human-AI teams [14, 192]. In human-AI collaboration, [24] show that incorporating models of human behavior into the training of the AI agent leads to higher collaborative performance compared to training the agent to play with themselves through self-play [162]. These studies highlight the importance of integrating human behavior into the design of collaborative AI. However, a key limitation of these research efforts is that they mostly assume that human behavior remains static, irrespective of the actions and behavior of the AI counterpart. In practice, humans may modify their behavior in response to their beliefs about what AI agents intend to do based on their observations of AI behavior.

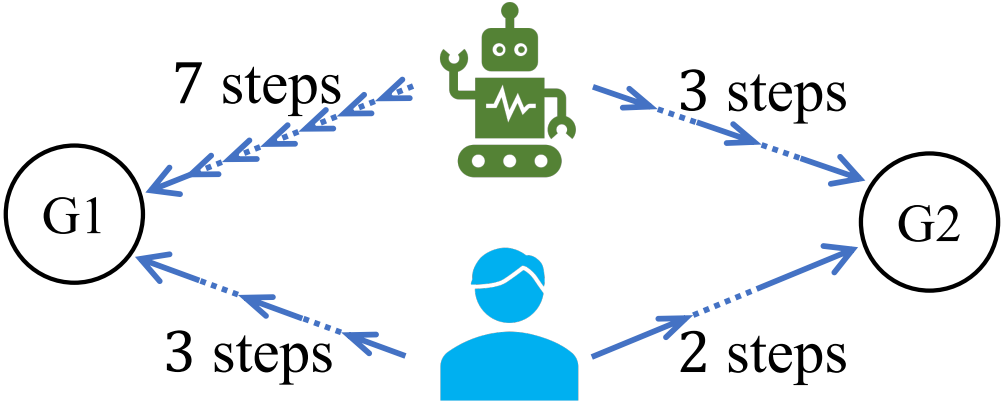


Figure 6.1: An example task of encoding beliefs about AI models into human-AI cooperation problem.

In this work, we argue that, in addition to incorporating human behavior, designing an AI agent that account for humans’ beliefs about AI behavior could significantly improve their collaborative performance. Consider a simple illustrative example in Figure 6.1, where the human and AI agents need to go to both goals $\{G1, G2\}$ to complete the task. Assume that the human prefers $G2$ over $G1$ because it requires fewer steps. When designing a collaborative AI that accounts for human behavior only, the AI would choose to go to $G1$, leading to lower overall collaborative performance. However, if the human would adjust their behavior based on their belief about AI behavior (e.g., they will avoid going to the same goal that the AI is going towards), we could incorporate this knowledge into the design of the AI. For example, the AI can choose to go to $G2$, anticipating that the human will account for that and go to $G1$, leading to improved overall performance. This example underscores the potential benefits of incorporating human beliefs about AI behavior in AI design. Meanwhile, it also

highlights the importance of developing accurate models of human beliefs and incorporating them into the AI design.

We formulate the human-AI collaboration environment as a multi-player goal-oriented MDP. We begin our investigation by developing models of human beliefs regarding AI behavior. This modeling effort extends the level- k reasoning framework [169] to account for suboptimal human behavior. Specifically, we first develop a *behavioral level-0* model that assumes agents take actions without considering the behavior of other agents. We then enhance the belief model by introducing a *behavioral level-1* model, which assumes humans interpret the behavior of another player as if the other agent adheres to the *behavioral level-0* model. With the models of human behavior and beliefs established, we proceed to develop collaborative AI agents that incorporate different assumptions about human models.

To examine proposed approaches, we conduct extensive human-subject experiments on two environments. To develop and assess the model of human behavior (i.e., the behavioral level-0 model), we first train models of human behavior using behavioral cloning on real-world human behavior and evaluate their performance. For the development and assessment of the model of human beliefs (i.e., the behavioral level-1 model), we conduct experiments that present each participant with a trace of behavior by an agent and ask them to infer the agent’s goal, examining whether our model leads to accurate predictions of human inference on the agent’s goal. Moreover, based on the belief model, we also explore the possibility of developing AI policies such that humans can more easily infer the goals of AI from its actions. Finally, utilizing the developed models of human behavior and beliefs, we conduct a final set of human-subject experiments that pair each human participant with an AI agent in a two-player coordination game. We assess the human-AI collaborative performance for different designs of AI agents. Overall, our results demonstrate the effectiveness of our developed model of human behavior and beliefs. Furthermore, we show that AI agents accounting for models of human behavior and beliefs achieve better collaborative performance with humans, compared to AI agents that do not consider human beliefs or those that disregard both human behavior and beliefs.

6.1 Related Work

This work joins the recent growing research in human-AI collaborations [24, 90, 14, 192]. [14] demonstrate that optimizing AI in isolation may lead to suboptimal performance for human-AI collaboration. [192] show that training AI systems to complement humans, performing better in areas where humans struggle, leads to improved collaborative performance. [24] illustrate that incorporating a human model, learned from human data, into the training of AI results in enhanced performance when these AI systems work with real humans. Additionally, [90] use population-based reinforcement learning to improve the robustness of trained AI agents. Our work extends this line of work by proposing to incorporate not only human behavior but also human beliefs of AI behavior into the design of collaborative AI.

From a technical perspective, our work involves understanding and modeling humans in decision-making. There has been a significant amount of work in the literature on modeling human behavior. For example, Inverse Reinforcement Learning (IRL)[129, 1, 146] aims to infer the reward functions in Markov Decision Processes (MDPs) through observing demonstrations of the optimal policy. If the demonstrator is a human being, the demonstrations could be noisy or contain behavioral biases. Studies [58, 159, 87, 215] have aimed to incorporate human behavioral biases in the inference process and infer both the rewards and biases simultaneously. Imitation Learning [88] also aims at developing models that can mimic human behavior from demonstrations. There have also been an increasing amount of research efforts that incorporates human models in computational and machine learning frameworks [100, 171, 172, 120, 121, 203, 205, 59].

From the perspective of modeling human beliefs over others' behavior, this has been discussed in level- k reasoning [169, 70] and theory of mind [142, 189]. As a few examples, [206] have found that real human behavior is close to level-1 and level-2 reasoning models in cooperative games. [4] survey works on autonomous agents modeling the beliefs and intentions of other agents. They distinguish methods for modeling stationary or changing agent behaviors. The belief model includes theory of mind, recursive reasoning models, plan recognition, partially observable Markov decision processes (POMDPs), and others. [10] propose a Bayesian theory of mind to describe human modeling of joint belief of state in a partially observable MDP and conduct human experiments in [12, 11]. A similar human modeling approach is adopted by [196] and applied to the Overcooked experiment. Their idea is to split delivering a dish into several subtasks, such as picking up an ingredient or putting an ingredient into a pot,

and the player is inferring others' subtasks and selecting their own corresponding subtask before taking actions.

6.2 Problem Formulations and Models

In this section, we first formulate the human-AI collaboration framework as a multi-player goal-oriented Markov decision process. We then outline our methods for modeling human behavior and beliefs about AI behavior. Finally, we present our approaches for integrating human behavior and beliefs into the development of collaborative AI agents.

6.2.1 Decision-Making Environment

We formulate the human-AI cooperative decision-making environment as a multi-agent MDP, as defined in Section 2.2. Note that in this work we consider the cooperative setting. Therefore, our formulation only incorporates a single reward function R for all players, though this can be easily extended.

While our formulation could address cases with multiple human and AI players, we focus on a two-player cooperative game in this Chapter, i.e., $\alpha = \{1, 2\}$, where one player is a human and the other is the AI. During the decision-making process, neither the human nor the AI knows the other player's next action or future plans, and they cannot communicate directly. However, they can observe each other's past actions, enabling them to infer about the other player and modify their own actions accordingly.

Goal-oriented MDP. In this work, we focus on the setting with goal-oriented MDP, where there is a set of goals $G = \{g_1, \dots, g_k\} \subseteq S$, which are subset of states that are terminal states, i.e., $P(g|g, a) = 1 \forall g \in G, a \in A$. Moreover, the decision-making agent only receives rewards when arriving at one of the goal states.

6.2.2 Modeling Human Behavior and Beliefs

Our models are motivated by the level- k framework [169] in economics. In particular, we start by considering humans as level-0 agents that do not account for others’ behavior in the environment. Differing from the literature, we address the natural situation that level-0 agents do not behave optimally, and we call this model *behavioral level-0* agents. To account for human beliefs,¹³ we model humans as an extension of the level-1 agents, that assume other agents are *behavioral* level-0 agents¹⁴ and update their beliefs in a Bayesian manner based on the observations of others’ behavior.

Modeling human behavior. We first model human behavior under the assumption that humans do not consider other players in the environment (or that they consider other players as a part of the environment without strategically responding). In this case, a human behavior model can be represented as $H : W \rightarrow \Pi$, mapping a given environment $w \in W$ to a policy $\pi = H(w)$. We give two examples of human behavior models utilized in our work below.

- *Standard model.* First consider the standard human behavior model in MDPs, in which the goal of the human is to maximize the expected cumulative reward, and their policy only depends on the current state. The model can be represented by $\pi(a|s)$, indicating the probability of choosing action a at state s . For the standard model that assumes decision optimality, humans choose actions maximizing the Q-function, where $Q(s, a)$ indicates the expected cumulative reward if the player takes action a in state s and follows policy π . $Q(s, a)$ could be calculated by standard reinforcement learning techniques such as value iteration or Q-learning in Section 2.3.
- *Behavioral level-0 model.* We also consider the case that we can learn human behavioral models from their historical behavioral traces through behavioral cloning [140, 178]. Behavioral cloning is one of the imitation learning approaches, which learns a policy from human demonstration by building a map from states to actions with supervised learning methods [9]. We build a fully connected neural network, where the input is the state encoding, and output is the probability over action space, and train the model with standard gradient descent method with cross-entropy loss.

¹³In this work, *human beliefs* refer to humans’ belief about the goal of the AI agent.

¹⁴Our model can iteratively progress to higher level- k agents. However, we focus on the case with $k \leq 1$.

Modeling human beliefs. We now describe how we develop models for human beliefs. As summarized earlier, we consider cases where human decision-makers assume that other agents in the environments are behavioral level-0 agents. In our discussion, we also describe this belief model as *behavioral level-1 agents*.

More specifically, we utilize Bayesian inference to model the human belief updating process, as in Equation (6.1), where $\lambda(g)$ represents the prior distribution of goals, and $Pr(s_t, a_t|g)$ denotes the probability of observing (s_t, a_t) given the goal, according to the policy model $\pi(a|s, g)$. This policy represents how humans perceive the actions of other players. If a human believes the other agent is following the standard model, then the policy is derived from the optimal policy, $\pi(a|s, g) \propto \exp(\beta Q(s, a|g))$ ¹⁵; if, however, humans believe the agent follows a data-driven model, then the policy will be the output of the human model, $\pi = H(w|g)$.

$$\begin{aligned}
 B(g|(s, a)_{1:t}) &\propto \lambda(g)Pr((s, a)_{1:t}|g) \\
 &= \lambda(g) \prod_{i=1}^{i=t} Pr(s_i, a_i|g) \\
 &= \lambda(g) \prod_{i=1}^{i=t} \pi(a_i|s_i, g)P(s_i|s_{i-1}, a_{i-1})
 \end{aligned} \tag{6.1}$$

6.2.3 Designing AI Agents

To illustrate the effectiveness of incorporating models of human beliefs, we train different AI agents that work with humans in human-AI collaboration problems via simulations and real-world human-subject experiments.

Training methodology. The main idea of our training method is through self-play. We incorporate the models of humans and have AI teammates play with the agents specified by the human models through simulated plays. We use PPO (as discussed in Section 2.3.3) to train collaborative AI agents.

¹⁵ $\beta \geq 0$ controls the level of optimality. When $\beta = \infty$, $\pi(a|s, g) = 1$ if $a = \operatorname{argmax}_{a'} Q(s, a'|g)$ and $\pi(a|s, g) = 0$ otherwise.

Collaborative AI agents. Using the methodology above, we have designed several AI agents based on different assumptions of the behavior of human counterparts.

- *Assuming humans are optimal.* We first train an AI agent that learns to collaborate with itself through self-play. This is equivalent to assuming the human counterpart is acting optimally.
- *Assuming humans are behavioral level-0 agents.* We next train an AI agent that assumes the human is a behavioral level-0 agent, using behavioral cloning to train the behavioral model.
- *Incorporating models of human behavior and beliefs.* Finally we also train an AI agent that incorporates both the models of human behavior and beliefs into the design of AI agents.

The details of the setup and implementations are included in Appendix B.4.

6.3 Experiments

We evaluate our approaches from multiple sets of experiments, consisting of both simulations and human-subject experiments in this Section. For our human-subject experiments, we have recruited in total 1,690 participants from Amazon Mechanical Turk (MTurk) for multiple sets of experiments. The experiments are approved by the IRB of our institution. Workers were paid \$1 base payments with the potential for bonus payments in some experiments. The average hourly rate is approximately \$14 across all our experiments.

6.3.1 Experiment Environments: Grid Worlds with Two Players

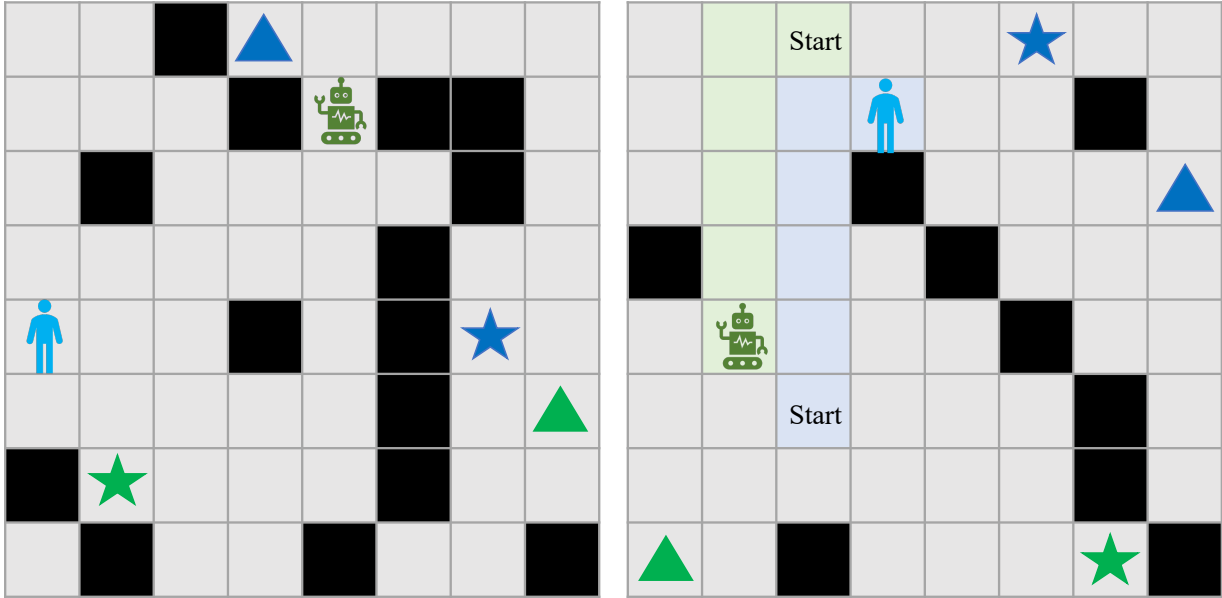
Our experiments are conducted in a grid world environment with two players and multiple goals. We have conducted two variations of the experiments. In the first variation, we designed the environments such that the two players are not playing in the same grid world. However, each of them has access to the full information of the environment and the actions of the other player. This variation abstracts away the interdependency of agent actions, meaning the agents' actions do not influence each other directly. This allows us to focus on how humans and AI reason about each other's goals and intentions. Note that we obtained

qualitatively similar results in this first variation and the second, more complicated variation of environments described next. For simplicity, we have included the setup and results of the first variation in the appendix. Below, we describe the experiment environment and results of our second variation.

More specifically, as shown in Figure 6.2, in our second variation of the environment, the grid world is 8 by 8 in size, containing the positions of both players and four possible goals. The players can choose to move {Up, Down, Right, Left} or stay in the current grid. They can see each other’s positions and take actions simultaneously. Each player can navigate to two of the four goals. In particular, the human player can only reach one of the two goals colored blue, and the AI player can only reach one of the goals colored green. When both agents reach the same type of goal (both reaching a "star" or both reaching a "triangle"), they earn positive points. However, they will not earn points if they reach different goals or if they collide into each other (move into the same position). We set the maximum number of actions to be 20.

In the first experiment, we recruit real-world humans to play against another pre-defined AI agent. We then examine the effectiveness of utilizing behavioral cloning as a model to mimic human behavior. Note that using behavioral cloning to model human behavior is extensively adopted in the literature, and the purpose of the first experiment serves as the foundation to develop our human belief models and the design of collaborative AI agents.

To evaluate our proposed approaches in modeling human beliefs and designing cooperative AI, we have designed and conducted additional two sets of experiments. In the second set of experiments, we provide participants with different traces of actions from other agents, and ask participants to infer the goal of the agents. This experiment helps us evaluate whether our belief model leads to better predictions of human beliefs over others’ behavior. Afterwards, in the third experiment, we conduct experiments similar to the first experiment. We team up AI agents and human players to examine the team performance of our design of collaborative AI when working with real-world human participants. But unlike the first experiment, we don’t reveal any information about goals the players are trying to reach.



(a) The interface for experiment 1 & 3.

(b) The interface for experiment 2.

Figure 6.2: The interface of human-subject experiments in human-AI cooperation. In Experiment 1, each participant is playing with an AI agent. The participants in experiment 1 are told what their goal is and only need to focus on reaching the goal without colliding with the AI agent. In Experiment 2, each participant is provided traces of the behavior by other agents and is asked to infer which goal one agent is trying to reach. Finally, in Experiment 3, the participants are not told which goal to reach, and they need to make decisions based on their beliefs over AI behavior. In Experiments 1 and 3, the participants can only receive bonus rewards by reaching the same type of goals (star or triangle) as the AI agent.

6.3.2 Experiment 1: Evaluating Human Behavior Models

In our first experiment, our goal is to examine the effectiveness of utilizing behavioral cloning to model human behavior. Our main purpose to conduct this experiment is to ensure whether this method works well in our setting, as our proposed approaches in developing human belief models and collaborative AI are built on top of models of human behavior.

Experiment setup. We recruited 190 workers from Amazon Mechanical Turk. Each recruited worker was asked to play 30 navigation games within a grid world with a pre-defined AI model, as shown in Figure 6.2a. To collect diverse data, we designed three AI models to play with humans: random AI (selecting a random goal and taking the shortest path), self-play AI (trained using self-play), and optimal AI (optimized via joint optimization). In Experiment 1, we provided additional instructions to participants about which goal they

should reach in each round, and the AI models controlled the other player to move towards the same type of goals. In our later Experiment 3, we did not provide these instructions, and humans needed to infer the goal of the AI agent.

Evaluation of data-driven behavior models. We divided the collected data of human actions into three sets: training, validation, and testing. The training set comprised data from 152 workers, including approximately 136,000 instances of user decisions, while the validation and testing sets each contained data from 19 workers, amounting to around 17,000 instances of user decisions each. We employed a 4-layer Multilayer Perceptron (MLP) model, where the input is the current environment layout, and the output predicts the next human action. We fine-tuned the hyperparameters, such as learning rate, hidden layer size, and L2 penalty, based on validation errors.

Our results suggest that human behavior is noisy and is significantly away from being optimal. Even when human participants are provided with suggested goals during data collection, there is only a 55.0% chance that both players will reach the same type of goals across all treatments. Besides, there is a notable chance of two players colliding (about 15.8%) or ending in different types of goals (about 25.6%). We compared the performance of our learned model to a model predicated on optimal agent behavior, defined as taking the shortest path to the goal. The training, validation, and test accuracies of both models are presented in Table 6.1. These results clearly reveal that human behavior deviates significantly from the assumed optimality. This deviation highlights the importance of incorporating a realistic model of human behavior into human-AI cooperation frameworks.

Table 6.1: The prediction accuracy for human behavior assuming optimal behavior and using data-driven model in Experiment 1.

	Training Accuracy	Validation Accuracy	Testing Accuracy
Assuming Optimal Behavior	0.4498	0.4327	0.4459
Data-Driven Model	0.8547	0.7831	0.7899

6.3.3 Experiment 2: Evaluating Human Belief Models

We next examine our model of human beliefs. Specifically, we investigate whether we can design AI behavior such that it is easier for humans to infer the goal of the AI agent.

In our experiments with a simplified environment, as included in Appendix B.4, we directly examined how accurately our belief model can infer human beliefs by comparing the predictions of our models with the direct solicitation of users’ inferences about others’ behavior. The results suggest that our model is more accurate in predicting human beliefs compared to baselines. However, human beliefs are generally very noisy, and if we randomly draw a behavior trace from another agent and ask humans to predict the goal, human prediction accuracy is close to random guessing. Therefore, we next shift our focus to whether AI agents can design their action plans to make it easier for humans to infer the goal. In this more complicated environment, we have directly examined this follow-up question of designing AI policies that are explicable to humans.

Experiment setup. The experiment setup is presented in Figure 6.2b. In addition to player positions and goal positions, we also display behavioral traces, which are sequences of actions taken by previous players. For each participant, we show them the traces of two players and ask the participant to infer which one of the two goals those agents are trying to reach.

We compare two belief models. The first one is the standard level-1 model: humans assume the other agent is a level-0 agent that takes the optimal decision (i.e., taking the shortest path towards the goal). The second one is our proposed *behavioral level-1* model: humans assume the other agent is also a suboptimal decision-maker, taking actions in the same way as themselves. In our implementation, we leverage the data-driven behavioral model derived from Experiment 1 as the user behavioral model. More concretely, we construct a belief model using Bayesian inference, as described in Section 6.2.2, using the corresponding human model of behavior.

Explicable AI policy. As mentioned earlier, humans generally struggle to infer the goals of other agents from behavioral traces randomly drawn from history. Therefore, in this experiment, instead of displaying a randomly drawn historical behavioral trace, we display the behavioral trace of an AI agent that aims to make its behavior *explicable*. In particular, based on the developed belief models, we train AI agents to not only achieve their goals but also to maximize the likelihood that humans can accurately infer these goals from their behavior. Our implementation utilizes self-play and awards an additional bonus when the goal inference aligns with the predictions of a human belief model, in addition to the rewards for goal achievement. The bonus is proportional to the log likelihood of belief model inference

$\log(\pi(a|s, g))$. By conducting experiments with this explicable AI policy, we simultaneously evaluate whether our belief models are accurate and whether our design of explicable AI policy is effective.

Simulations. We design the AI agents to incorporate the belief models of standard level-1 agents and behavioral level-1 agents, enabling them to adopt policies that simplify the task of goal inference for humans. For each environment with a given goal, we generated the behavioral traces of both AI policies. The goal is to examine whether the explicable AI with our belief model makes it easier for humans to infer the goal of the AI. We first conducted simulations assuming humans are behavioral level-1 agent. The results are shown in Figure 6.3a. The explicable AI equipped with the belief model of behavioral level-1 agents leads to behavioral traces that are easier to infer. Note that this result is not surprising since we assume humans infer the goal following the beliefs of behavioral level-1 agent. The results provide evidence that our design of explicable AI is effective.

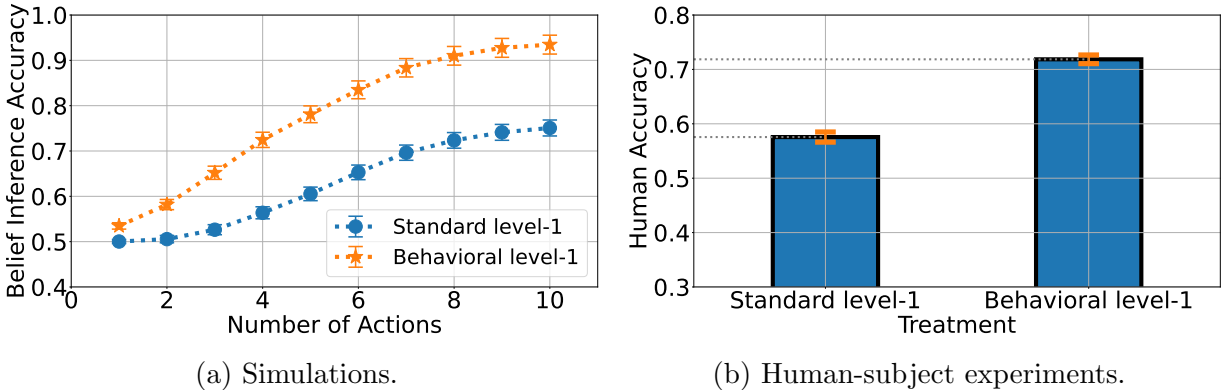


Figure 6.3: Belief inference results in simulations and human-subject experiments in Experiment 2.

Human-subject experiments. We next conducted human-subject experiments to assess whether our approach results in actions that make it easier for real humans to infer the goal. We recruited 200 workers from Amazon Mechanical Turk, randomly assigning them to one of two treatments. Each participant was tasked with identifying the goals of the player in 30 different scenarios. The length of actions is drawn from the range 4 to 8. To incentivize effort, participants were awarded a \$0.03 bonus for each correctly identified goal. The results, as shown in Figure 6.3b, demonstrate that humans are better at inferring the goal of the explicable AI agent, that assumes humans are behavioral level-1 agents. The difference is statistical significant with $p > 0.0001$. This comparison demonstrates that when

coupled with the more accurate human belief models, we can indeed design AI policy to be explicable, i.e., making it easier for humans to infer AI goals from observing AI actions.

6.3.4 Experiment 3: Designing Collaborative AI Agents

We now examine our design of collaborative AI agents. As described in Section 6.2.3, we train collaborative AI agents using three different corresponding human models and utilize both simulations and human-subject experiments to examine the effectiveness of our design.

Experiment setup. The setup of our Experiment 3 is the same as Experiment 1, shown in Figure 6.2a. But human participants will not receive any hint about which goal their teammate or themselves are suggested to reach.

Simulations. We first run simulations to evaluate the performance of human-AI teams with different design of collaborative AI. The evaluation is based on 10,000 randomly generated environments. However, we further filter out cases where the distance between the two goals for the same player is smaller than 3 to encourage models to adjust their behavior based on the inference of their teammate’s actions. We examine the performance of three designs of collaborative AI using the methodology described in Section 6.2.3:

- *Self-Play* is the AI that is trained assuming they are playing with itself.
- *Behavior-AI* is the AI that assumes the human partner is the behavioral level-0 agent.
- *Behavior&Belief-AI* is the AI that assumes the human partner is the behavioral level-1 agent.

We partner the collaborative AI agents with the three simulated human agents and measure the collaborative performance. The simulation results, shown in Table 6.2, highlight the importance of selecting appropriate human models for training collaborative AI. When an AI agent is paired with the human model that is used to train the AI, the collaborative performance is better compared to pairing with other models.

Human-subject experiments. To examine the performance of our collaborative AI design when pairing with real humans, we recruited 200 workers and divided them into three treatment groups, each interacting with a different design of AI. Participants were tasked with playing 30 games, preceded by three tutorial games designed to familiarize them with

Table 6.2: Simulation results of collaborative performance over 10k testing cases in Experiment 3. Column players are different AI agents, and row players are different simulated human models.

Human Model	AI Agent		
	Self-play	Behavior-AI	Behavior&Belief-AI
Self-play AI	0.6245	0.5250	0.6177
Behavior Model	0.4902	0.6334	0.5755
Behavior&Belief	0.6411	0.6574	0.7675

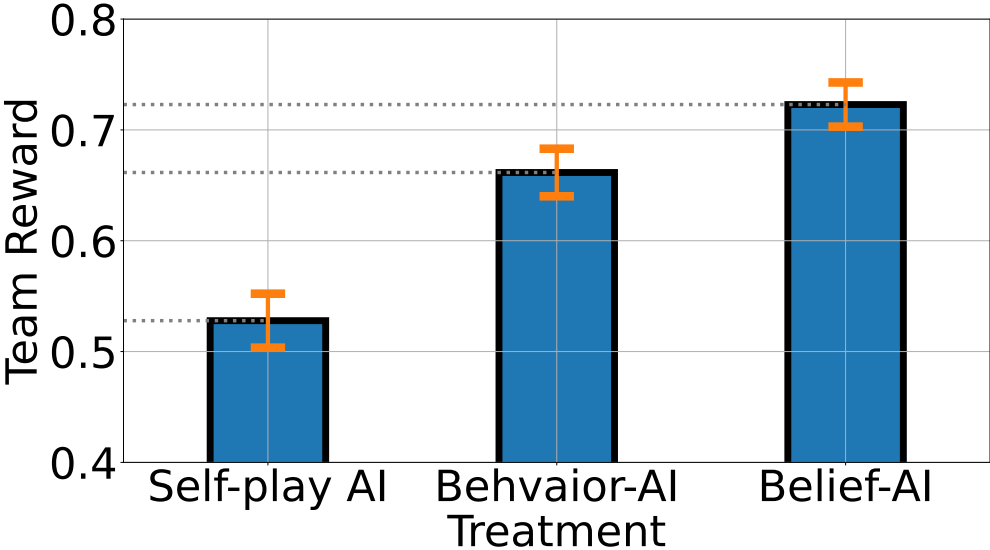


Figure 6.4: The average human-AI collaboration performance in human-subject experiment in Experiment 3.

the gameplay. Participants can receive a \$0.05 bonus payment for reaching the same goal with the AI agent for each of the 30 tasks.

Figure 6.4 presents the average collaborative reward, highlighting that AI trained with models of human behavior and beliefs achieved the highest collaborative performance when working with real humans. This outcome not only validates our design of AI but also suggests that real humans’ actions align with the behavior and belief models in this environment. Statistical analysis shows significant differences in the performance of AI models paired with humans, with $p < 0.0001$ for comparisons between self-play and AI accounting for human behaviors, and a p value of 0.0014 for comparisons between AI accounting for human behaviors versus accounting for both behavior and belief. These results demonstrate that incorporating

human beliefs into the design of AI agents enhances collaborative performance when working with real humans.

6.4 Discussions

This work explores the impact of incorporating human behavior and beliefs into the design of collaborative AI, aiming to improve collaboration between humans and AI. By developing models that account for human beliefs regarding AI actions and integrating these models into AI design, we have observed improved performance in human-AI collaboration. Our approach, validated through simulations and human-subject experiments, demonstrates that AI agents designed with an understanding of human behavior and beliefs can be more effective in working with humans. This suggests a potential path forward for creating AI systems that are better aligned with human partners, making collaborative tasks more efficient and intuitive.

Generalization and limitations. We have demonstrated the effectiveness of our approach through human-subject experiments in commonly used grid world environments with varying levels of complexity. However, as with most human-subject studies, our findings are limited to the chosen environments. Moreover, our approach leverages historical data to construct models of human behavior and beliefs. This approach implicitly assumes that both human behavior and beliefs remain unchanged over time. Therefore, it is important to improve the development of human models and examine their generalizability to other environments.

Societal impacts. Our approach highlights the importance of incorporating not only human behavior but also human beliefs about AI behavior in the design of AI for human-AI collaboration. We believe this will become increasingly important as AI capabilities grow. However, while our research focuses on enhancing human-AI collaboration, there is a potential for the same methods to be used negatively, such as designing AI that intentionally sabotages human utility. One approach to mitigate potential negative impacts is to increase the transparency of AI models, enabling humans to develop appropriate reliance when working with AI agents.

Chapter 7

Conclusion and Future Directions

In summary, this dissertation investigates how to model human decision-makers and design AI systems for effective interaction with real humans. We show that, in various one-shot and sequential decision-making problems, real humans often deviate from the common assumption of rationality. Ignoring these biases poses significant challenges in designing AI systems, impacting both their performance and computational efficiency.

In order to influence human decision-makers, we begin by modifying the decision-making environments in Chapter 3. We incorporate time-inconsistent bias models into the environment design problem and demonstrate that solving such problems is NP-hard. We propose two approaches: modifying reward functions and sending real-time nudges, and show their effectiveness in a navigation game involving human participants. Although these bias models align well with real humans in our experimental setup (where humans have limited visibility of the environment), this method is constrained by the specific assumptions about human behavior models. To address this limitation, we extend our method-based approach to a data-driven approach in Chapter 4, where AI systems design information policies to influence human decision-makers. The data-driven approach does not rely on assumptions about human biases; instead, it requires sufficient data collected from real humans. We find that this approach can more accurately predict human behaviors. Furthermore, we extend our work beyond maximizing a pre-defined utility function to consider how to elicit human ethical preferences and how AI systems can generate predictive information to assist humans. In Chapter 5, we demonstrate that AI can generate predictive information to influence decision-making in kidney transplant scenarios, showing that humans are more likely to accept suggestions from an AI system that shares similar ethical preferences. In Chapter 6, we further explore scenarios where AI systems make joint decisions with human decision-makers, aiming to maximize team performance in cooperative tasks. Real-time cooperation introduces new challenges, as human players may adjust their behavior to better

collaborate with AI teammates. We use a data-driven approach to model human behavior in a two-player navigation problem and develop a belief model to describe humans' beliefs about their AI teammates. Incorporating these belief models allows us to train AI systems to cooperate more effectively with real humans.

Overall, our proposed framework leverages both method-based and data-driven approaches to model human behavior and beliefs about AI systems, enabling the design of AI systems that interact effectively with humans by updating information signals, modifying decision-making environments, and developing AI teammates. While we demonstrate the effectiveness of our proposed approaches, there are several limitations, including challenges in data collection from crowdsourcing platforms, accurately modeling human behavior, and addressing potential ethical concerns.

Challenges in data collection. We conducted our human-subject experiments and evaluated the designed AI systems on Amazon Mechanical Turk. Due to the distributed nature of crowd work, we cannot guarantee that workers are sufficiently engaged with the tasks, which might partially explain their deviation from rational or optimal behavior. Common approaches to improve the quality of crowdsourced data collection include assigning tasks to suitable workers [83, 80], performing post-hoc aggregation [39, 191, 79, 214], designing proper incentives [119, 78, 85, 84, 199, 82, 81], and improving task design [62, 3, 44, 176, 46, 47]. From an experimental design perspective, we added bonuses proportional to performance in some experiments to encourage careful decision-making. Additionally, we included extra tasks in some experiments to check the consistency of human responses and later used this as a filter to remove potential noisy responses. Despite these efforts, we cannot guarantee the high quality of the collected dataset. In particular, in Chapter 5, we study ethical decision-making, and due to the subjective nature of these tasks, it is difficult to evaluate whether workers are providing truthful answers. Moreover, we surveyed the ethical preferences of a general population of laypeople, who might have different interpretations of moral dilemmas (e.g., whether they believe another kidney will be available soon). It might be helpful to survey the preferences of relevant domain stakeholders. For example, in the domain of kidney transplants, we might want to elicit preferences from medical doctors or policymakers. In the domain of autonomous vehicles, preferences could be gathered from car manufacturers, drivers, or pedestrians.

Challenges in modeling human decision-makers. Our work demonstrates that data-driven methods, such as supervised learning and behavioral cloning, can predict human actions more accurately than assuming humans are rational. However, we also find that real human behavior and beliefs are often noisy and inconsistent. A successful AI system relies on an accurate human model, which in turn depends heavily on the availability of sufficient high-quality human data. Efficiently building or collecting this data remains a crucial future direction. Ensuring the data quality and representativeness is essential to improve the reliability of human behavior models. In the context of sequential decision-making problems, even if our human models can predict individual actions accurately, the overall task-level accuracy—such as recovering the entire decision path—often falls short. This discrepancy highlights the complexity of human behavior in more extended and intricate tasks. The challenge is not just predicting single actions but understanding and anticipating the series of actions that comprise human decision-making over time. Moreover, modeling human behavior in complex tasks introduces additional layers of difficulty. Human decision-making in real-world scenarios is influenced by various factors, including emotions, stress, fatigue, and unforeseen events. Capturing these elements are beyond the capacity of this dissertation, but necessary for creating AI systems that can interact effectively with humans in realistic settings.

Comparisons between method-based and data-driven approaches. We utilize both method-based and data-driven approaches to model human decision-makers. Method-based approaches, as discussed in Chapter 3 and Chapter 4, rely on prior knowledge about human behavior in specific tasks. These approaches can lead to analytical solutions for AI systems, clearly revealing the effects of biases. However, they are limited by their reliance on predefined assumptions about human behavior, and might be hard to transferred to new scenarios. On the other hand, data-driven approaches, highlighted in Chapter 4 and Chapter 6, generally offer more accurate predictions given sufficient data and do not require extra assumptions. These models, however, present challenges in AI system design because optimal solutions may not translate well when the agent is represented by a neural network. They require large amounts of high-quality data, are more complex to design, and can be less interpretable, making it harder to find optimal solutions compared to method-based approaches. Despite this, we employ data-driven approach (deep learning) to design AI systems as it does not depend on the formulation of the decision-making model. For realistic problems, hybrid approaches, combining both method-based and data-driven approaches,

might be a good choice. Balancing these approaches based on the specific requirements and constraints of the task at hand can lead to more effective AI system design.

Future work in personalization. The data-driven approach builds behavior models from datasets collected from crowdsourcing workers. Though this helps to collect diverse data, it also makes it difficult for us to create a personalized model. Personalization has shown its superiority in many AI systems, such as recommender systems. In applications where we can collect massive data for a single person, such as autonomous driving or behavior tracking on mobile phones, it's possible to utilize personalized data to develop a more differentiated model. One possible approach is through training models on multiple datasets and then fine-tuning them for customization. Another way is to encode personalized data into embeddings and represent individuals' preferences or biases as part of AI systems. However, personalization raises privacy concerns during data collection and model deployment. Many existing research efforts have been conducted in this area, but the advancement of AI will likely bring more challenges regarding privacy.

Future work in human-AI alignment. In most of our work, we often assume a predefined reward function for both human decision-makers and AI models, whether aligned or misaligned. However, in practice, rewards might not be directly given in many applications, making it challenging to ensure that AI behaves in ways that align with the interests of the designer. Additionally, we find that collecting human beliefs is more difficult than collecting human behavior. This issue becomes more pronounced in applications involving ethical decision-making (as discussed in Chapter 5) or in rapidly changing domains such as social media moderation and financial markets. Ethical decision-making often involves diverse and sometimes conflicting ethical standards. AI systems must navigate these complexities to make decisions that are broadly acceptable and ethically sound. Moreover, ethical principles can evolve over time, influenced by cultural shifts, societal debates, and legal changes. Therefore, AI systems need mechanisms to update their ethical frameworks accordingly. Addressing these challenges is crucial for the future development of AI systems that are aligned with human interests.

References

- [1] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *International Conference on Machine Learning*, 2004.
- [2] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [3] H. Alagarai Sampath, R. Rajeshuni, and B. Indurkha. Cognitively inspired task design to improve user performance on crowdsourcing platforms. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3665–3674, 2014.
- [4] S. V. Albrecht and P. Stone. Autonomous agents modelling other agents: A comprehensive survey and open problems. *Artificial Intelligence*, 258:66–95, 2018.
- [5] R. Alonso and O. Câmara. Persuading voters. *American Economic Review*, 106(11):3590–3605, 2016.
- [6] J. Angwin, J. Larson, S. Mattu, and L. Kirchner. Machine bias. *ProPublica*, May, 23(2016):139–159, 2016.
- [7] R. Apel, I. Erev, R. Reichart, and M. Tennenholtz. Predicting decisions in language based persuasion games. *Journal of Artificial Intelligence Research*, 73:1025–1091, 2022.
- [8] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J.-F. Bonnefon, and I. Rahwan. The moral machine experiment. *Nature*, 2019.
- [9] M. Bain and C. Sammut. A framework for behavioural cloning. In *Proceedings of Machine Intelligence 15*, pages 103–129, 1995.
- [10] C. Baker, R. Saxe, and J. Tenenbaum. Bayesian theory of mind: Modeling joint belief-desire attribution. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 33, 2011.
- [11] C. L. Baker, J. Jara-Ettinger, R. Saxe, and J. B. Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4):1–10, 2017.
- [12] C. L. Baker, R. Saxe, and J. B. Tenenbaum. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009.

- [13] A. V. Banerjee. A simple model of herd behavior. *The quarterly journal of economics*, 107(3):797–817, 1992.
- [14] G. Bansal, B. Nushi, E. Kamar, E. Horvitz, and D. S. Weld. Is the most accurate AI the best teammate? optimizing AI for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11405–11414, 2021.
- [15] G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz. Beyond accuracy: The role of mental models in human-ai team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, volume 7, pages 2–11, 2019.
- [16] A. Bartik and S. Nelson. Credit reports as resumes: The incidence of pre-employment credit screening. 2016.
- [17] A. Bauer, D. Wollherr, and M. Buss. Human–robot collaboration: a survey. *International Journal of Humanoid Robotics*, 5(01):47–66, 2008.
- [18] D. Bergemann and S. Morris. Information design: A unified perspective. *Journal of Economic Literature*, 57(1):44–95, 2019.
- [19] R. Berk. An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *Journal of Experimental Criminology*, 13(2):193–216, 2017.
- [20] Y. E. Bigman, D. Wilson, M. N. Arnestad, A. Waytz, and K. Gray. Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General*, 2022.
- [21] R. Binns. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 514–524, 2020.
- [22] D. D. Bourgin, J. C. Peterson, D. Reichman, S. J. Russell, and T. L. Griffiths. Cognitive model priors for predicting human decisions. In *International Conference on Machine Learning*, pages 5133–5141, 2019.
- [23] P. Bracke, A. Datta, C. Jung, and S. Sen. Machine learning explainability in finance: an application to default risk analysis. 2019.
- [24] M. Carroll, R. Shah, M. K. Ho, T. Griffiths, S. Seshia, P. Abbeel, and A. Dragan. On the utility of learning about humans for human-AI coordination. In *Proceedings of the Advances in Neural Information Processing Systems*, 2019.
- [25] M. Castiglioni, A. Celli, A. Marchesi, and N. Gatti. Online bayesian persuasion. *Advances in Neural Information Processing Systems*, 33:16188–16198, 2020.

- [26] L. Chan, A. Critch, and A. Dragan. Human irrationality: both bad and good for reward inference. *arXiv preprint arXiv:2111.06956*, 2021.
- [27] L. Chan, K. Doyle, D. McElfresh, V. Conitzer, J. P. Dickerson, J. Schaich Borg, and W. Sinnott-Armstrong. Artificial artificial intelligence: Measuring influence of ai’assessments’ on moral decision-making. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 214–220, 2020.
- [28] J. Chen, M. Li, and H. Xu. Selling data to a machine learner: Pricing via costly signaling. In *International Conference on Machine Learning*, pages 3336–3359, 2022.
- [29] Y. Cheng, F. Wang, P. Zhang, and J. Hu. Risk prediction with electronic health records: A deep learning approach. In *Proceedings of the SIAM International Conference on Data Mining*, pages 432–440, 2016.
- [30] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems*, 29, 2016.
- [31] R. Choudhury, G. Swamy, D. Hadfield-Menell, and A. D. Dragan. On the utility of model learning in hri. In *ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 317–325, 2019.
- [32] A. Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.
- [33] R. Cole and T. Roughgarden. The sample complexity of revenue maximization. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 243–252, 2014.
- [34] V. Conitzer and T. Sandholm. Complexity of mechanism design. In *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, pages 103–110, 2002.
- [35] S. Corbett-Davies and S. Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*, 2018.
- [36] D. Cornelisse, T. Rood, Y. Bachrach, M. Malinowski, and T. Kachman. Neural payoff machines: Predicting fair and stable payoff allocations among team members. In *Advances in Neural Information Processing Systems*, 2022.
- [37] M. Curry, P.-Y. Chiang, T. Goldstein, and J. Dickerson. Certifying strategyproof auction networks. In *Advances in Neural Information Processing Systems*, pages 4987–4998, 2020.
- [38] G. de Clippel and X. Zhang. Non-bayesian persuasion. Technical report, Technical report, Working Paper, 2019.

- [39] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38, 1977.
- [40] B. J. Dietvorst, J. P. Simmons, and C. Massey. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114, 2015.
- [41] J. J. Dijkstra, W. B. Liebrand, and E. Timminga. Persuasiveness of expert systems. *Behaviour & Information Technology*, 17(3):155–163, 1998.
- [42] B. Ding, Y. Feng, C.-J. Ho, W. Tang, and H. Xu. Competitive information design for pandora’s box. In *Proceedings of the 2023 Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 353–381, 2023.
- [43] Z. Dong, S. Das, P. Fowler, and C.-J. Ho. Efficient nonmyopic online allocation of scarce reusable resources. In *AAMAS Conference proceedings*, 2021.
- [44] R. Drapeau, L. Chilton, J. Bragg, and D. Weld. Microtalk: Using argumentation to improve crowdsourcing accuracy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 4, 2016.
- [45] I. E. Dror and D. Charlton. Why experts make errors. *Journal of Forensic Identification*, 56(4):600, 2006.
- [46] X. Duan, C.-J. Ho, and M. Yin. Does exposure to diverse perspectives mitigate biases in crowdwork? an explorative study. In *AAAI Conference on Human Computation and Crowdsourcing*, 2020.
- [47] X. Duan, C.-J. Ho, and M. Yin. The influences of task design on crowdsourced judgement: A case study of recidivism risk evaluation. In *The Web Conference (WWW)*, 2022.
- [48] S. Dughmi and H. Xu. Algorithmic bayesian persuasion. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pages 412–425, 2016.
- [49] S. Dughmi and H. Xu. Algorithmic persuasion with no externalities. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 351–368, 2017.
- [50] P. Dütting, Z. Feng, H. Narasimhan, D. Parkes, and S. S. Ravindranath. Optimal auctions through deep learning. In *International Conference on Machine Learning*, pages 1706–1715, 2019.
- [51] C. Dwork. Differential privacy. In *International Colloquium on Automata, Languages, and Programming*, pages 1–12, 2006.

- [52] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference*, pages 265–284, 2006.
- [53] E. J. Emanuel, G. Persad, R. Upshur, B. Thome, M. Parker, A. Glickman, C. Zhang, C. Boyle, M. Smith, and J. P. Phillips. Fair allocation of scarce medical resources in the time of covid-19. *New England Journal of Medicine*, 382(21):2049–2055, 2020.
- [54] E. J. Emanuel and A. Wertheimer. Who should get influenza vaccine when not all can? *Science*, 312(5775):854–855, 2006.
- [55] E. J. Emanuel and A. Wertheimer. Who should get influenza vaccine when not all can? *Science*, 312(5775):854–855, 2006.
- [56] Y. Emek, M. Feldman, I. Gamzu, R. PaesLeme, and M. Tennenholtz. Signaling schemes for revenue maximization. *ACM Transactions on Economics and Computation (TEAC)*, 2(2):1–19, 2014.
- [57] O. Evans and N. D. Goodman. Learning the preferences of bounded agents. In *NIPS Workshop on Bounded Optimality*, 2015.
- [58] O. Evans, A. Stuhlmüller, and N. Goodman. Learning the preferences of ignorant, inconsistent agents. In *AAAI Conference on Artificial Intelligence*, 2016.
- [59] Y. Feng, C.-J. Ho, and W. Tang. Rationality-robust information design: Bayesian persuasion under quantal response. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*, 2024.
- [60] Y. Feng, W. Tang, and H. Xu. Online bayesian recommendation with no regret. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 818–819, 2022.
- [61] Z. Feng, H. Narasimhan, and D. C. Parkes. Deep learning for revenue-optimal auctions with budgets. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems*, pages 354–362, 2018.
- [62] A. Finnerty, P. Kucherbaev, S. Tranquillini, and G. Convertino. Keep it simple: Reward and task design in crowdsourcing. In *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI*, pages 1–4, 2013.
- [63] P. Frazier, D. Kempe, J. Kleinberg, and R. Kleinberg. Incentivizing exploration. In *ACM Conference on Economics and Computation*, 2014.
- [64] R. Freedman, J. S. Borg, W. Sinnott-Armstrong, J. P. Dickerson, and V. Conitzer. Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence*, 283:103261, 2020.

- [65] R. Freedman, J. Schaich Borg, W. Sinnott-Armstrong, J. P. Dickerson, and V. Conitzer. Adapting a kidney exchange algorithm to align with human values. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, page 115, New York, NY, USA, 2018.
- [66] A. Furnham. Factors relating to the allocation of medical resources. *Journal of Social Behavior and Personality*, 11(3):615–624, 1996.
- [67] A. Furnham. Factors relating to the allocation of medical resources. *Journal of Social Behavior and Personality*, 11(3):615–624, 1996.
- [68] A. Fuster, P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther. Predictably unequal? the effects of machine learning on credit markets. *The Journal of Finance*, 77(1):5–47, 2022.
- [69] S. Gehlbach and K. Sonin. Government control of the media. *Journal of public Economics*, 118:163–171, 2014.
- [70] D. Gill and V. Prowse. Cognitive ability, character skills, and learning to play equilibrium: A level-k analysis. *Journal of Political Economy*, 124(6):1619–1676, 2016.
- [71] I. Goldstein and Y. Leitner. Stress tests and information disclosure. *Journal of Economic Theory*, 177:34–69, 2018.
- [72] N. Golowich, H. Narasimhan, and D. C. Parkes. Deep learning for multi-facility location mechanism design. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 261–267, 2018.
- [73] S. Gottwald and D. A. Braun. Bounded rational decision-making from elementary computations that reduce uncertainty. *Entropy*, 2019.
- [74] U. Gretzel and D. R. Fesenmaier. Persuasion in recommender systems. *International Journal of Electronic Commerce*, 11(2):81–100, 2006.
- [75] N. Grgić-Hlača, C. Castelluccia, and K. P. Gummadi. Taking advice from (dis) similar machines: The impact of human-machine similarity on machine-assisted decision-making. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 10, pages 74–88, 2022.
- [76] N. Grgic-Hlaca, E. M. Redmiles, K. P. Gummadi, and A. Weller. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference*, page 903–912, 2018.
- [77] K. Harris, V. Chen, J. Kim, A. Talwalkar, H. Heidari, and S. Z. Wu. Bayesian persuasion for algorithmic recourse. *Advances in Neural Information Processing Systems*, 35:11131–11144, 2022.

- [78] C.-J. Ho and K.-T. Chen. On formal models for social verification. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pages 62–69, 2009.
- [79] C.-J. Ho, R. Frongillo, and Y. Chen. Eliciting categorical data for optimal aggregation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 2450–2458, 2016.
- [80] C.-J. Ho, S. Jabbari, and J. W. Vaughan. Adaptive task assignment for crowdsourced classification.. 2013. In *Proc. 30th Int. Conf. on Machine Learning*, 2013.
- [81] C.-J. Ho, A. Slivkins, S. Suri, and J. W. Vaughan. Incentivizing high quality crowd-work. In *Proceedings of the 24th International Conference on World Wide Web*, pages 419–429, 2015.
- [82] C.-J. Ho, A. Slivkins, and J. W. Vaughan. Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 359–376, 2014.
- [83] C.-J. Ho and J. Vaughan. Online task assignment in crowdsourcing markets. In *Proceedings of the AAAI conference on artificial intelligence*, volume 26, pages 45–51, 2012.
- [84] C.-J. Ho, Y. Zhang, J. W. Vaughan, and M. Van Der Schaar. Towards social norm design for crowdsourcing markets. In *Proceedings of the 4th Human Computation Workshop*, 2012.
- [85] J. J. Horton and L. B. Chilton. The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM conference on Electronic commerce (EC)*, 2010.
- [86] A. Hudon, T. Demazure, A. Karran, P.-M. Léger, and S. Sénécal. Explainable artificial intelligence (xai): how the visualization of ai predictions affects user cognitive load and confidence. In *Information Systems and Neuroscience: NeuroIS Retreat 2021*, pages 237–246. Springer, 2021.
- [87] D. Hughes, A. Agarwal, Y. Guo, and K. Sycara. Inferring non-stationary human preferences for human-agent teams. In *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2020.
- [88] A. Hussein, M. M. Gaber, E. Elyan, and C. Jayne. Imitation learning: A survey of learning methods. *ACM Computing Surveys*, 50(2):1–35, 2017.
- [89] B. Ibarz, J. Leike, T. Pohlen, G. Irving, S. Legg, and D. Amodei. Reward learning from human preferences and demonstrations in atari. *Advances in neural information processing systems*, 31, 2018.

- [90] M. Jaderberg, W. M. Czarnecki, I. Dunning, L. Marris, G. Lever, A. G. Castaneda, C. Beattie, N. C. Rabinowitz, A. S. Morcos, A. Ruderman, et al. Human-level performance in 3d multiplayer games with population-based reinforcement learning. *Science*, 364(6443):859–865, 2019.
- [91] D. Kahneman. A perspective on judgment and choice: mapping bounded rationality. *American Psychologist*, 58(9):697, 2003.
- [92] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pages 99–127. World Scientific, 2013.
- [93] E. Kamenica. Bayesian persuasion and information design. *Annual Review of Economics*, 11:249–272, 2019.
- [94] E. Kamenica and M. Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.
- [95] R. Kasumba, G. Yu, C.-J. Ho, S. Keren, and W. Yeoh. Data-driven goal recognition design for general behavioral agents. *arXiv preprint arXiv:2404.03054*, 2024.
- [96] B. F. Klare, M. J. Burge, J. C. Klontz, R. W. V. Bruegge, and A. K. Jain. Face recognition performance: Role of demographic information. *IEEE Transactions on Information Forensics and Security*, 7(6):1789–1801, 2012.
- [97] J. Klayman and Y.-W. Ha. Confirmation, disconfirmation, and information in hypothesis testing. *Psychological review*, 94(2):211, 1987.
- [98] J. Kleinberg, J. Ludwig, S. Mullainathan, and C. R. Sunstein. Discrimination in the age of algorithms. *Journal of Legal Analysis*, 10, 2018.
- [99] J. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.
- [100] J. Kleinberg and S. Oren. Time-inconsistent planning: a computational problem in behavioral economics. In *ACM Conference on Economics and Computation*, 2014.
- [101] J. Kleinberg, S. Oren, and M. Raghavan. Planning with multiple biases. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 567–584, 2017.
- [102] P. Krütli, T. Rosemann, K. Y. Törnblom, and T. Smieszek. How to fairly allocate scarce medical resources: ethical argumentation under scrutiny by health professionals and lay people. *PloS one*, 11(7), 2016.
- [103] K. Kuo, A. Ostuni, E. Horishny, M. J. Curry, S. Dooley, P.-y. Chiang, T. Goldstein, and J. P. Dickerson. Proportionnet: Balancing fairness and revenue for auction design with deep learning. *arXiv preprint arXiv:2010.06398*, 2020.

- [104] M. Kwon, E. Biyik, A. Talati, K. Bhasin, D. P. Losey, and D. Sadigh. When humans aren't optimal: Robots that collaborate with risk-aware humans. In *ACM/IEEE International Conference on Human-Robot Interaction*, pages 43–52, 2020.
- [105] M. K. Lee, D. Kusbit, A. Kahng, J. T. Kim, X. Yuan, A. Chan, D. See, R. Noothigattu, S. Lee, A. Psomas, and A. D. Procaccia. Webuildai: Participatory framework for algorithmic governance. *Proc. ACM Hum.-Comput. Interact.*, 3, 2019.
- [106] S. Lewandowsky, M. Mundy, and G. Tan. The dynamics of trust: comparing humans to automation. *Journal of Experimental Psychology: Applied*, 6(2):104, 2000.
- [107] Z. Li, K. Lieberman, W. Macke, S. Carrillo, C.-J. Ho, J. Wellen, and S. Das. Incorporating compatible pairs in kidney exchange: A dynamic weighted matching model. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 349–367, 2019.
- [108] D. Lingenbrink and K. Iyer. Optimal signaling mechanisms in unobservable queues. *Operations research*, 67(5):1397–1416, 2019.
- [109] X. Liu, C. Yu, Z. Zhang, Z. Zheng, Y. Rong, H. Lv, D. Huo, Y. Wang, D. Chen, J. Xu, et al. Neural auction: End-to-end learning of auction mechanisms for e-commerce advertising. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3354–3364, 2021.
- [110] Y. Liu and C.-J. Ho. Incentivizing high quality user contributions: New arm generation in bandit learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [111] J. M. Logg. Theory of machine: When do people rely on algorithms? *Harvard Business School working paper series# 17-086*, 2017.
- [112] J. M. Logg, J. A. Minson, and D. A. Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103, 2019.
- [113] C. Longoni, A. Bonezzi, and C. K. Morewedge. Resistance to medical artificial intelligence. *Journal of Consumer Research*, 46(4):629–650, 2019.
- [114] X. Luo, A. Recharadt, G. Sun, K. K. Nejad, F. Yáñez, B. Yilmaz, K. Lee, A. O. Cohen, V. Borghesani, A. Pashkov, et al. Large language models surpass human experts in predicting neuroscience results. *arXiv preprint arXiv:2403.03230*, 2024.
- [115] G. Mallapragada, S. R. Chandukala, and Q. Liu. Exploring the effects of “what”(product) and “where”(website) characteristics on online shopping behavior. *Journal of Marketing*, 80(2):21–38, 2016.

- [116] Y. Mansour, A. Slivkins, and V. Syrgkanis. Bayesian incentive-compatible bandit exploration. In *ACM Conference on Economics and Computation*, 2015.
- [117] E. Mark, D. Goldsman, B. Gurbaxani, P. Keskinocak, and J. Sokol. Using machine learning and an ensemble of methods to predict kidney transplant survival. *PloS one*, 14(1), 2019.
- [118] W. Mason and S. Suri. Conducting behavioral research on amazon’s mechanical turk. *Behavior research methods*, 44(1):1–23, 2012.
- [119] W. Mason and D. Watts. Financial incentives and the “performance of crowds”. In *Proceedings of the 1st Human Computation Workshop (HCOMP)*, 2009.
- [120] P. Masters, M. Kirley, and W. Smith. Extended goal recognition: a planning-based model for strategic deception. In *International Conference on Autonomous Agents and MultiAgent Systems*, 2021.
- [121] P. Masters, W. Smith, and M. Kirley. Extended goal recognition: Lessons from magic. *Frontiers in Artificial Intelligence*, 4, 2021.
- [122] D. McCloskey and A. Klamer. One quarter of gdp is persuasion. *The American Economic Review*, 85(2):191–195, 1995.
- [123] D. McFadden. Econometric models of probabilistic choice. *Structural Analysis of Discrete Data with Econometric Applications*, 198272, 1981.
- [124] D. McFadden. Economic choices. *American economic review*, 91(3):351–378, 2001.
- [125] S. Mehrotra, C. M. Jonker, and M. L. Tielman. More similar values, more trust?-the effect of value similarity on trust in human-agent interaction. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 777–783, 2021.
- [126] P. Mobadersany, S. Yousefi, M. Amgad, D. A. Gutman, J. S. Barnholtz-Sloan, J. E. Velázquez Vega, D. J. Brat, and L. A. Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13):E2970–E2979, 2018.
- [127] S. Narayanan, G. Yu, C.-J. Ho, and M. Yin. How does value similarity affect human reliance in ai-assisted ethical decision making? In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 2023.
- [128] S. Narayanan, G. Yu, W. Tang, C.-J. Ho, and M. Yin. How does predictive information affect human ethical preferences? In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022.
- [129] A. Y. Ng, S. J. Russell, et al. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*, 2000.

- [130] R. S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220, 1998.
- [131] J. Nilsson, M. Ohlsson, P. Höglund, B. Ekmehag, B. Koul, and B. Andersson. The international heart transplant survival algorithm (ihtsa): a new model to improve organ sharing and survival. *PloS one*, 10(3), 2015.
- [132] R. Noothigattu, S. Gaikwad, E. Awad, S. Dsouza, I. Rahwan, P. Ravikumar, and A. Procaccia. A voting-based system for ethical decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [133] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.
- [134] T. O’Donoghue and M. Rabin. Doing it now or later. *American Economic Review*, 89(1):103–124, 1999.
- [135] N. Peri, M. Curry, S. Dooley, and J. Dickerson. Preferencenet: Encoding human preferences in auction design with deep learning. *Advances in Neural Information Processing Systems*, 34:17532–17542, 2021.
- [136] G. Persad, A. Wertheimer, and E. J. Emanuel. Principles for allocation of scarce medical interventions. *The lancet*, 373(9661):423–431, 2009.
- [137] G. Persad, A. Wertheimer, and E. J. Emanuel. Principles for allocation of scarce medical interventions. *The Lancet*, 373(9661):423–431, 2009.
- [138] J. C. Peterson, D. D. Bourgin, M. Agrawal, D. Reichman, and T. L. Griffiths. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, 372(6547):1209–1214, 2021.
- [139] G. M. Piscitello, E. M. Kapania, W. D. Miller, J. C. Rojas, M. Siegler, and W. F. Parker. Variation in ventilator allocation guidelines by us state during the coronavirus disease 2019 pandemic: a systematic review. *JAMA network open*, 3(6), 2020.
- [140] D. A. Pomerleau. Alvin: An autonomous land vehicle in a neural network. In *Proceedings of the Advances in Neural Information Processing Systems*, 1988.
- [141] D. Prelec. The probability weighting function. *Econometrica*, pages 497–527, 1998.
- [142] D. Premack and G. Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, 1978.
- [143] M. Promberger and J. Baron. Do patients trust computers? *Journal of Behavioral Decision Making*, 19(5):455–468, 2006.

- [144] J. Rahme, S. Jelassi, and S. M. Weinberg. Auction learning as a two-player game. In *International Conference on Learning Representations*, 2020.
- [145] A. Rakhsha, G. Radanovic, R. Devidze, X. Zhu, and A. Singla. Policy teaching via environment poisoning: Training-time adversarial attacks against reinforcement learning. In *International Conference on Machine Learning*, 2020.
- [146] D. Ramachandran and E. Amir. Bayesian inverse reinforcement learning. In *International Joint Conference on Artificial Intelligence*, 2007.
- [147] L. Rayo and I. Segal. Optimal information disclosure. *Journal of Political Economy*, 118(5):949–987, 2010.
- [148] S. Reddy, S. Levine, and A. Dragan. Assisted perception: optimizing observations to communicate state. In *Conference on Robot Learning*, pages 748–764. PMLR, 2021.
- [149] M. O. Rieger and M. Wang. Cumulative prospect theory and the st. petersburg paradox. *Economic Theory*, 28(3):665–679, 2006.
- [150] A. Robinson, S. Booker, and K. Gauntt. Eliminate use of dsa and region from kidney allocation one year post-implementation monitoring report, 2023.
- [151] S. J. Rosenbaum et al. Ethical considerations for decision making regarding allocation of mechanical ventilators during a severe influenza pandemic or other public health emergency. 2011.
- [152] T. Sandholm. Automated mechanism design: A new application area for search algorithms. In *International Conference on Principles and Practice of Constraint Programming*, pages 19–36. Springer, 2003.
- [153] T. Sandholm and A. Likhodedov. Automated design of revenue-maximizing combinatorial auctions. *Operations Research*, 63(5):1000–1025, 2015.
- [154] N. A. Saxena, K. Huang, E. DeFilippis, G. Radanovic, D. C. Parkes, and Y. Liu. How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 99–106, 2019.
- [155] S. Schach, S. Gottwald, and D. A. Braun. Quantifying motor task performance by bounded rational decision theory. *Frontiers in neuroscience*, 2018.
- [156] A. Schmitt, T. Wambsganss, M. Söllner, and A. Janson. Towards a trust reliance paradox? exploring the gap between perceived trust in and reliance on algorithmic advice. In *International Conference on Information Systems (ICIS)*, 2021.
- [157] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

- [158] V. A. Shaffer, C. A. Probst, E. C. Merkle, H. R. Arkes, and M. A. Medow. Why do patients derogate physicians who use a computer-based diagnostic support system? *Medical Decision Making*, 33(1):108–118, 2013.
- [159] R. Shah, N. Gundotra, P. Abbeel, and A. Dragan. On the feasibility of learning, rather than assuming, human biases for reward inference. In *International Conference on Machine Learning*, 2019.
- [160] F. Sherwani, M. M. Asad, and B. S. K. K. Ibrahim. Collaborative robots and industrial revolution 4.0 (IR 4.0). In *Proceedings of the International Conference on Emerging Trends in Smart Technologies*, pages 1–5, 2020.
- [161] M. Siegrist, G. Cvetkovich, and C. Roth. Salient value similarity, social trust, and risk/benefit perception. *Risk analysis*, 20(3):353–362, 2000.
- [162] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.
- [163] S. B. Sitkin and N. L. Roth. Explaining the limited effectiveness of legalistic “remedies” for trust/distrust. *Organization science*, 4(3):367–392, 1993.
- [164] A. Slivkins. Introduction to multi-armed bandits. *Found. Trends Mach. Learn.*, 2019.
- [165] K. A. Small. A discrete choice model for ordered alternatives. *Econometrica: Journal of the Econometric Society*, pages 409–424, 1987.
- [166] C. E. Smith, B. Yu, A. Srivastava, A. Halfaker, L. Terveen, and H. Zhu. Keeping community in the loop: Understanding wikipedia stakeholder values for machine learning-based systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.
- [167] Z. Song, R. Parr, and L. Carin. Revisiting the softmax bellman operator: New benefits and new perspective. In *International Conference on Machine Learning*, 2019.
- [168] M. Srivastava, H. Heidari, and A. Krause. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 2459–2468, 2019.
- [169] D. O. Stahl II and P. W. Wilson. Experimental evidence on players’ models of other players. *Journal of Economic Behavior & Organization*, 25(3):309–327, 1994.
- [170] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

- [171] W. Tang and C.-J. Ho. Bandit learning with biased human feedback. In *International Conference on Autonomous Agents and Multiagent Systems*, 2019.
- [172] W. Tang and C.-J. Ho. On the bayesian rational assumption in information design. In *AAAI Conference on Human Computation and Crowdsourcing*, 2021.
- [173] W. Tang, C.-J. Ho, and Y. Liu. Differentially private contextual dynamic pricing. In *International Conference on Autonomous Agents and Multiagent Systems*, pages 1368–1376, 2020.
- [174] W. Tang, C.-J. Ho, and Y. Liu. Bandit learning with delayed impact of actions. *Advances in Neural Information Processing Systems*, 34:26804–26817, 2021.
- [175] W. Tang, C.-J. Ho, and Y. Liu. Linear models are robust optimal under strategic behavior. In *International Conference on Artificial Intelligence and Statistics*, pages 2584–2592. PMLR, 2021.
- [176] W. Tang, C.-J. Ho, and M. Yin. Leveraging peer communication to enhance crowd-sourcing. In *The Web Conference (WWW)*, 2019.
- [177] S. Tolmeijer, M. Christen, S. Kandul, M. Kneer, and A. Bernstein. Capable but amoral? comparing ai and human expert collaboration in ethical decision making. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–17, 2022.
- [178] F. Torabi, G. Warnell, and P. Stone. Behavioral cloning from observation. *arXiv preprint arXiv:1805.01954*, 2018.
- [179] K. E. Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.
- [180] L. Treiman, C.-J. Ho, and W. Kool. Humans forgo reward to instill fairness into AI. In *AAAI Conference on Human Computation and Crowdsourcing*, 2023.
- [181] L. Treiman, C.-J. Ho, and W. Kool. The consequences of AI training on human decision making. *Proceedings of the National Academy of Sciences (PNAS)*, 2024.
- [182] B. Ustun and C. Rudin. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3):349–391, 2016.
- [183] N. Van Berkel, J. Goncalves, D. Russo, S. Hosio, and M. B. Skov. Effect of information presentation on fairness perceptions of machine learning predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2021.
- [184] E. Vayena, A. Blasimme, and I. G. Cohen. Machine learning in medicine: addressing ethical challenges. *PLoS medicine*, 15(11), 2018.

- [185] N. Vellodi. Ratings design and barriers to entry. *Available at SSRN 3267061*, 2018.
- [186] J. Von Neumann and O. Morgenstern. Theory of games and economic behavior: 60th anniversary commemorative edition. In *Theory of games and economic behavior*. Princeton university press, 2007.
- [187] R. Wang, F. M. Harper, and H. Zhu. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020.
- [188] J. Wei, C. Yang, X. Song, Y. Lu, N. Hu, D. Tran, D. Peng, R. Liu, D. Huang, C. Du, et al. Long-form factuality in large language models. *arXiv preprint arXiv:2403.18802*, 2024.
- [189] S. Westby and C. Riedl. Collective intelligence in human-AI teams: A Bayesian theory of mind approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6119–6127, 2023.
- [190] D. B. White, M. H. Katz, J. M. Luce, and B. Lo. Who should receive life support during a public health emergency? using ethical principles to improve allocation decisions. *Annals of Internal Medicine*, 150(2):132–138, 2009.
- [191] J. Whitehill, T. fan Wu, J. Bergsma, J. R. Movellan, and P. 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.
- [192] B. Wilder, E. Horvitz, and E. Kamar. Learning to complement humans. In *Proceedings of the International Conference on International Joint Conferences on Artificial Intelligence*, pages 1526–1533, 2021.
- [193] H. J. Wilson and P. R. Daugherty. Collaborative intelligence: Humans and AI are joining forces. *Harvard Business Review*, 96(4):114–123, 2018.
- [194] J. R. Wolf. Do it students prefer doctors who use it? *Computers in Human Behavior*, 35:287–294, 2014.
- [195] G. Wu and R. Gonzalez. Curvature of the probability weighting function. *Management Science*, 42(12):1676–1690, 1996.
- [196] S. A. Wu, R. E. Wang, J. A. Evans, J. B. Tenenbaum, D. C. Parkes, and M. Kleiman-Weiner. Too many cooks: Bayesian inference for coordinating multi-agent collaboration. *Topics in Cognitive Science*, 13(2):414–432, 2021.
- [197] H. Xu. On the tractability of public persuasion with no externalities. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2708–2727, 2020.

- [198] H. Xu, K. Wang, P. Vayanos, and M. Tambe. Strategic coordination of human patrollers and mobile sensors with signaling for security games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.
- [199] M. Yin and Y. Chen. Predicting crowd work quality under monetary interventions. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 4, pages 259–268, 2016.
- [200] M. Yin, J. Wortman Vaughan, and H. 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.
- [201] R. Yokoi and K. Nakayachi. The effect of value similarity on trust in the automation systems: A case of transportation and medical care. *International Journal of Human-Computer Interaction*, 37(13):1269–1282, 2021.
- [202] B. Yu, Y. Yuan, L. Terveen, Z. S. Wu, J. Forlizzi, and H. Zhu. *Keeping Designers in the Loop: Communicating Inherent Algorithmic Trade-Offs Across Multiple Objectives*, page 1245–1257. Association for Computing Machinery, New York, NY, USA, 2020.
- [203] G. Yu and C.-J. Ho. Environment design for biased decision makers. In *IJCAI*, pages 592–598, 2022.
- [204] G. Yu, R. Kasumba, C.-J. Ho, and W. Yeoh. On the utility of accounting for human beliefs about ai behavior in human-ai collaboration, 2024.
- [205] G. Yu, W. Tang, S. Narayanan, and C.-J. Ho. Encoding human behavior in information design through deep learning. *Advances in Neural Information Processing Systems*, 36, 2024.
- [206] J. Yu, G. Pressoir, W. H. Briggs, I. Vroh Bi, M. Yamasaki, J. F. Doebley, M. D. McMullen, B. S. Gaut, D. M. Nielsen, J. B. Holland, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38(2):203–208, 2006.
- [207] E. N. Zalta, U. Nodelman, C. Allen, et al. Stanford encyclopedia of philosophy, 1995.
- [208] H. Zhang, Y. Chen, and D. C. Parkes. A general approach to environment design with one agent. In *International Joint Conference on Artificial Intelligence*, 2009.
- [209] H. Zhang and D. C. Parkes. Value-based policy teaching with active indirect elicitation. In *AAAI Conference on Artificial Intelligence*, 2008.
- [210] H. Zhang, J. Wang, Z. Zhou, W. Zhang, Y. Wen, Y. Yu, and W. Li. Learning to design games: Strategic environments in reinforcement learning. In *International Joint Conference on Artificial Intelligence*, 2018.

- [211] S. Zhang and A. J. Yu. Forgetful bayes and myopic planning: Human learning and decision-making in a bandit setting. In *Advances in Neural Information Processing Systems*, 2013.
- [212] X. Zhang, Y. Ma, A. Singla, and X. Zhu. Adaptive reward-poisoning attacks against reinforcement learning. In *International Conference on Machine Learning*, 2020.
- [213] Y. Zhang, Q. V. Liao, and R. K. Bellamy. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 295–305, 2020.
- [214] Y. Zheng, G. Li, Y. Li, C. Shan, and R. Cheng. Truth inference in crowdsourcing: Is the problem solved? *Proceedings of the VLDB Endowment*, 10(5):541–552, 2017.
- [215] T. Zhi-Xuan, J. Mann, T. Silver, J. Tenenbaum, and V. Mansinghka. Online bayesian goal inference for boundedly rational planning agents. In *Advances in Neural Information Processing Systems*, 2020.

Appendix A

Proofs of Theorems

We provide proofs of our theorems in the main body of the dissertation.

A.1 Proof of Lemma 1

Proof. We prove the lemma by constructing an example MDP for bounded-rational agents. Consider a bounded-rational agent with parameter τ . We construct an MDP as show in Figure A.1, where the circle denotes the state, arrow denotes the action (with deterministic transition), and the number associate with the arrow is the reward. We consider the case that the reward functions for the principal and the agent are the same. In this example, the set of state is $\{s_0, \dots, s_{\tau+1}\}$. For states s_i with $i = 1$ to τ , there is only one available action “move right” that moves to state s_{i+1} , where the reward $R^a(s_i, \text{move right}) = R^p(s_i, \text{move right}) = 1$. For state s_0 , there is an additional action of staying in state s_0 that lead to reward of 2, and for state $s_{\tau+1}$, the only action is to move to state s_0 that gives a reward of m .

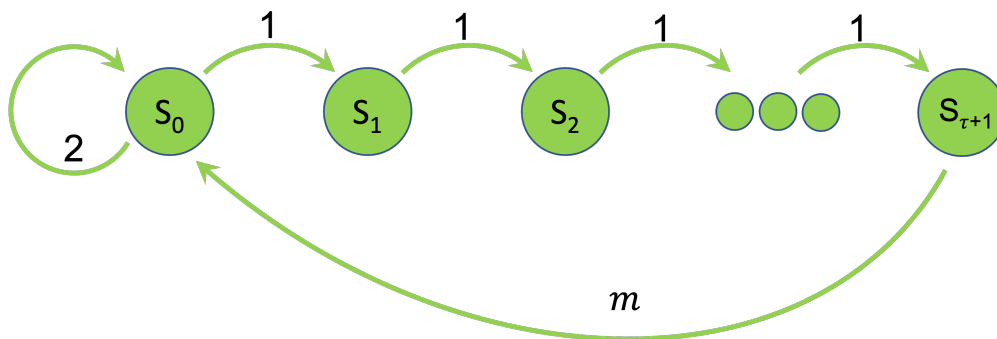


Figure A.1: The example MDP used for proving Lemma 1 with bounded-rational agents.

Let the initial state be s_0 and $T = \tau + 2$. If the agent is bounded rational with τ , it is easy to see that the agent will choose to stay in s_0 and generate a total reward of $(\tau + 2) * 2$ for the principal. If the agent is a standard agent with $\gamma = 1$, he will move from s_0 to $s_{\tau+1}$ then back to s_0 , leading to a total reward of $\tau + 1 + m$. Therefore, without environment design, the ratio between the reward by a bounded-rational agent and the reward by a standard agent would be $2(\tau + 2)/(\tau + 1 + m)$, which goes to 0 when $m \rightarrow \infty$. This proves that there exists an MDP such that the ratio of the reward made by an bounded rational agent compared to the reward made by a standard agent is arbitrarily close to 0.

If the principal believes the agent is a standard agent, she does not need to update the environment to reach the optimal payoff. However, the agent, who is bounded rational, would take the the sub-optimal action. Therefore, the principal’s performance ratio would again be $2(\tau + 2)/(\tau + 1 + m)$, which could goes to 0 when $m \rightarrow \infty$. \square

A.2 Proof of Theorem 2

Proof. We prove the NP-hardness through a reduction from the knapsack problem. The reduction process is similar to the one by [209], though we use the myopic agent case to show that the problem is NP-hard with biased agents. Note that this proof holds for both the problem of environment design via reward function modification and via action nudge.

An instance of the knapsack problem consists of n items. Denote the i -th item value as $u_i > 0$ and the weight as $w_i > 0$, for $1 \leq i \leq n$. The knapsack problem aims to find a set of items that maximizes the total values while ensuring the total weights is within budget B . Using variables $x_i \in \{0, 1\}$ to denote whether item i is included in the set, the knapsack problem can be formulated as the below integer program, which is NP-hard in general:

$$\max_x \sum_{i=1}^n u_i x_i; \text{ s.t. } \sum_{i=1}^n w_i x_i \leq B, \quad x_i \in \{0, 1\}, \forall i \quad (\text{A.1})$$

For the reduction, for each instance of the knapsack problem, we can construct an MDP as follows. Consider an MDP with $n + 1$ states, s_1 to s_{n+1} . The initial state is s_1 . We also make the action set from state s_{n+1} to be empty, effectively making it an end state. There are two actions $\{a_1, a_0\}$ available for each state, except for s_{n+1} , with taking a_1 at state s_i representing accepting item i , and taking a_0 representing not accepting. No matter which action is taken at state s_i , the next state will always be s_{i+1} . The agent’s reward function corresponds to the weight for accepting item and the principal’s rewards corresponds to the utility, i.e., agent reward is $R^a(s_i, a_1) = -w_i$, and principal reward is $R^p(s_i, a_1) = u_i$. Taking action a_0 leads to zero reward for both principal and agent, i.e., $R^a(s_i, a_0) = R^p(s_i, a_0) = 0$. Let the agent be myopic, and the budget for the principal is B . Note that in this MDP, the initial state is s_1 , and no matter what the agent policy is, the state at time t is s_t , so we have omitted the time index in the agent policy to simplify the presentation, which means the myopic agent

will choose $\pi(s_i) = \operatorname{argmax}_{a \in \{a_0, a_1\}} R^a(s_i, a)$. Note also that, without environment design, the myopic agent will always choose action a_0 since $R^a(s_i, a_0) = 0 > -w_i = R^a(s_i, a_1), \forall i$, and therefore the principal's total reward is 0 in this case.

The environment design problem either via reward function modification or action nudge can be expressed in the same way as in Equation (A.2). The reason is that 1) in this MDP construction, the agent policy does not depend on time t , and 2) since the agent is myopic, both reward function modification and action nudge need to pay same amount of $c(s_i, a)$ for agent to change action. Therefore, the optimization formulation for both design spaces is the same as in Equation (A.2), and this NP-hardness proof holds for both cases.

$$\begin{aligned}
& \max_c \sum_{i=1}^n R^p(s_i, \pi(s_i)) \\
& \text{s.t.} \sum_{i=1}^n \sum_{a \in \{a_0, a_1\}} |c(s_i, a)| \leq B \\
& \pi(s_i) = \operatorname{argmax}_a \{R^a(s_i, a) + c(s_i, a)\}, \forall s_i
\end{aligned} \tag{A.2}$$

Below we show that, with the solution of Equation (A.2), we can obtain the solution of the knapsack problems in polynomial time. Since we can construct the environment design problem for every instance of the knapsack problem, if our environment design problem is not NP-hard, the knapsack problem is not NP-hard, which lead to the contradiction since the knapsack problem is known to be NP-hard. Observe that if we have the solution $c(s, a)$ from Equation (A.2), we can obtain $\pi(s_i)$ as well in the equality constraint. If $\pi(s_i) = a_0$, we have $c(s_i, a_0) = c(s_i, a_1) = 0$. If $\pi(s_i) = a_1$, we have $c(s_i, a_1) - c(s_i, a_0) \geq R^a(s_i, a_0) - R^a(s_i, a_1) = w_i$. Therefore, $|c(s_i, a_1)| + |c(s_i, a_0)| \geq w_i$, and the equality holds when we set $c(s_i, a_1) = w_i$ and $c(s_i, a_0) = 0$. With the above observation and our MDP construction, setting x_i to be 1 in the knapsack problem if and only if $\pi(s_i) = a_1$ could maximize the total utility of selected items while satisfying the budget constraint on item weights. This means we can solve the knapsack problem if the solution of Equation (A.2) is given. This finishes the proof. \square

A.3 Proof of Lemma 3

Recall that we define $Q(s, a, t)$ as $Q(s, a, t, 0)$. Below we give the proof of a more general version of Lemma 3, as stated below.

Lemma 5. *For any environment w , let π_w and ρ_w be the agent's deterministic and stochastic policies following our model. Let $Q^{\pi_w}(s, a, t, \hat{t})$ and $Q^{\rho_w}(s, a, t, \hat{t})$ be the corresponding Q -functions. For all (s, a, t, \hat{t}) , we have*

$$|Q^{\pi_w}(s, a, t, \hat{t}) - Q^{\rho_w}(s, a, t, \hat{t})| \leq \mathcal{O}(e^{-\beta C}),$$

where $C > 0$ is a constant and β is the parameter of ρ .

Proof. This proof extends the results by [167], who prove the convergence for infinite-time horizon MDP, to address finite horizon and general discounting function. In the following proof, we omit the subscript w in π_w and ρ_w and represent them using π and ρ . For a biased agent with discounting function $d(t)$, Let $Q^\pi(s, a, t, \hat{t})$ be the biased Q -function following π . Similar to the standard notation convention, we use the random variable s_t^π to denote state at time t when following policy π . The expectation is taken over the randomness of state transition. Also, since we are considering finite-horizon MDP, we set $Q^\pi(s, a, t, \hat{t}) = 0$ for $t + \hat{t} > T$, which represents the unreachable horizon. We have

$$Q^\pi(s, a, t, \hat{t}) = d(\hat{t})R(s, a) + \mathbb{E}[Q^\pi(s_{t+\hat{t}+1}^\pi, \pi(s_{t+\hat{t}+1}^\pi, t, \hat{t} + 1), t, \hat{t} + 1)]$$

We can write down $Q^\rho(s, a, t, \hat{t})$ similarly as Q^π . The only difference is in the second term. Instead of taking action $\pi(s_{t+\hat{t}+1}^\pi, t, \hat{t} + 1)$, the agent takes action $a = a_{t+\hat{t}+1}^\rho$ with probability $\rho(s, a, t, \hat{t} + 1) = \frac{e^{\beta Q^\rho(s, a, t, \hat{t} + 1)}}{\sum_{a'} e^{\beta Q^\rho(s, a', t, \hat{t} + 1)}}$. Moreover, the expectation is taken over both state transition and stochastic policy.

$$Q^\rho(s, a, t, \hat{t}) = d(\hat{t})R(s, a) + \mathbb{E}[Q^\rho(s_{t+\hat{t}+1}^\rho, a_{t+\hat{t}+1}^\rho, t, \hat{t} + 1)]$$

Claim 1: $Q^\pi(s, a, t, \hat{t}) - Q^\rho(s, a, t, \hat{t}) \geq 0$

Proof. We prove this claim by induction. Note that by definition, both Q functions are 0 when $\hat{t} + t > T$.

- When $\hat{t} = T - t$, $Q^\pi(s, a, t, T - t) - Q^\rho(s, a, t, T - t) = R(s, a) - R(s, a) = 0$.
- When $\hat{t} < T - t$, we have $Q^\pi(s, a, t, \hat{t} - 1) - Q^\rho(s, a, t, \hat{t} - 1) = \mathbb{E}[\max_a Q^\pi(s, a, t, \hat{t}) - \sum_a \rho(s, a, t, \hat{t}) Q^\rho(s, a, t, \hat{t})]$. Since $Q^\pi(s, a, t, \hat{t}) \geq Q^\rho(s, a, t, \hat{t})$ for all (s, a, t, \hat{t}) , we have $Q^\pi(s, a, t, \hat{t}) - Q^\rho(s, a, t, \hat{t}) \geq 0$.

□

For the purpose of the analysis, we define two functions:

$$\begin{aligned} \delta(s, t, \hat{t}) &= \max_a Q^\pi(s, a, t, \hat{t}) - \sum_a \rho(s, a, t, \hat{t}) Q^\pi(s, a, t, \hat{t}) \\ \zeta(t, \hat{t}) &= \max_s \delta(s, t, \hat{t}) \end{aligned}$$

Claim 2: $Q^\pi(s, a, t, \hat{t}) - Q^\rho(s, a, t, \hat{t}) \leq \sum_{j=t+\hat{t}+1}^T \zeta(t, j - t)$

Proof. We again prove it by induction.

- When $\hat{t} = T - t$, $Q^\pi(s, a, t, T - t) - Q^\rho(s, a, t, T - t) = R(s, a) - R(s, a) = 0$.
- Suppose the statement is true for \hat{t} . For $\hat{t} - 1$, we have (for notation simplicity, we use s' to denote $s_{t+\tau}$). The expectation of s' is over the state transition $P(s'|s, a)$ and the expectation of a' is over the stochastic policy ρ .

$$\begin{aligned}
& Q^\pi(s, a, t, \hat{t} - 1) - Q^\rho(s, a, t, \hat{t} - 1) \\
&= \mathbb{E} \left[\max_{s'} Q^\pi(s', a', t, \hat{t}) - \mathbb{E}_{a'}[Q^\rho(s', a', t, \hat{t})] \right] \\
&\leq \mathbb{E} \left[\max_{s'} Q^\pi(s', a', t, \hat{t}) - \mathbb{E}_{a'}[Q^\pi(s', a', t, \hat{t})] + \sum_{j=t+\hat{t}+1}^T \zeta(t, j - t) \right] \\
&= \mathbb{E}[\delta(s', t, \hat{t})] + \sum_{j=t+\hat{t}+1}^T \zeta(t, j - t) \\
&\leq \zeta(t, \hat{t}) + \sum_{j=t+\hat{t}+1}^T \zeta(t, j - t) \\
&= \sum_{j=t+\hat{t}}^T \zeta(t, j - t)
\end{aligned}$$

Therefore, by induction, we know the claim is true. \square

With the above claims, we now show how ζ converges in terms of β . For given (s, t, \hat{t}) , we sort $\{Q^\pi(s, a, t, \hat{t})\}$ such that $Q^\pi(s, a_{[1]}, t, \hat{t}) \geq Q^\pi(s, a_{[2]}, t, \hat{t}) \geq \dots \geq Q^\pi(s, a_{[m]}, t, \hat{t})$. Therefore, we have $\sigma_i = Q^\pi(s, a_{[1]}, t, \hat{t}) - Q^\pi(s, a_{[i]}, t, \hat{t}) \geq 0$. Also, there exists an index $i^* \leq m$ such that $\sigma_i > 0, \forall i^* \leq i \leq m$ and $\sigma_i = 0, \forall i < i^*$. If i^* does not exist, for all action $Q^\pi(s, a, t, \hat{t}) = \max_{a'} Q^\pi(s, a', t, \hat{t})$, there is no difference in selecting any action. Notice that i^* depends on (s, t, \hat{t}) , but we omit the dependency for clarity.

Note that we can express $\delta(s, t, \hat{t})$ as below.

$$\begin{aligned}
\delta(s, t, \hat{t}) &= Q^\pi(s, a_{[1]}, t, \hat{t}) - \sum_a \rho(s, a, t, \hat{t}) Q^\pi(s, a, t, \hat{t}) \\
&= \frac{\sum_{i=2}^m e^{-\beta\sigma_i} \sigma_i}{1 + \sum_{i=2}^m e^{-\beta\sigma_i}}
\end{aligned}$$

Since $\frac{\sum_i x_i}{1 + \sum_i y_i} \leq \sum_i \frac{x_i}{1 + y_i}$ for non-negative sequences $\{x_i\}$ and $\{y_i\}$. By setting $x_i = e^{-\beta\sigma_i} \sigma_i$ and $y_i = e^{-\beta\sigma_i}$, we have

$$\begin{aligned}
\delta(s, t, \hat{t}) &= \frac{\sum_{i=2}^m e^{-\beta\sigma_i} \sigma_i}{1 + \sum_{i=2}^m e^{-\beta\sigma_i}} \\
&\leq \sum_{i=2}^m \frac{e^{-\beta\sigma_i} \sigma_i}{1 + e^{-\beta\sigma_i}} \\
&= \sum_{i=2}^m \frac{\sigma_i}{1 + e^{\beta\sigma_i}} \\
&= \sum_{i=i^*}^m \frac{\sigma_i}{1 + e^{\beta\sigma_i}} \\
&\leq e^{-\beta\sigma_{i^*}} \sum_{i=i^*}^m \sigma_i
\end{aligned}$$

Therefore, $\zeta(t, \hat{t})$ can be upper bounded as below:

$$\zeta(t, \hat{t}) = \max_s \delta(s, t, \hat{t}) \leq \max_s e^{-\beta\sigma_{i^*}} \sum_{i=i^*}^m \sigma_i$$

By applying Claim 2, we have the following

$$\begin{aligned}
&Q^\pi(s, a, t, \hat{t}) - Q^\rho(s, a, t, \hat{t}) \\
&\leq \sum_{j=t+\hat{t}+1}^T \zeta(t, j-t) \\
&\leq (T-t-\hat{t}) \max_{t+\hat{t}+1 \leq j \leq T} \max_s e^{-\beta\sigma_{i^*}} \sum_{i=i^*}^m \sigma_i
\end{aligned}$$

Note that $\sigma_i \leq \max_{s,a,t,\hat{t}} Q^\pi(s, a, t, \hat{t}) \leq \sum_{t=0}^T d(t) R_{max}$, and $R_{max} = \max_{s,a} R(s, a) > 0$. If we choose $\sigma^* = \min_{i,s,t,\hat{t}} \sigma_i$ such that $\sigma_i > 0$ holds, the following bounds holds for all (s, a, t, \hat{t}) and $\beta > \beta_0$:

$$Q^\pi(s, a, t, \hat{t}) - Q^\rho(s, a, t, \hat{t}) \leq [m(T+1) \sum_{t=0}^T d(t) R_{max}] e^{-\beta\sigma^*}$$

which finishes our proof. □

A.4 Proof of Lemma 4

Proof. The problem formulated in Equation (3.6) is a LP problem with $|S|(T+1) + 1$ constraints, with $|S|T$ constraints on transition dynamics, $|S|$ constraints on initial distribution, and 1 constraint on the nudge budget (excluding the constraints $\phi(s, a, t) \geq 0$). Therefore, using the property of linear programs, there exists at least one optimal solution with at most $|S|(T+1) + 1$ non-zero variables (the one with the smallest number of non-zero variables is called the basic feasible solution).

First consider a special case that we can find an optimal solution ϕ^* that (1) has at most $|S|(T+1) + 1$ non-zero variables and (2) there exists an action a for every (s, t) such that $\phi^*(s, a, t) > 0$. Note that finding ϕ^* satisfying (1) is always possible using the property of linear programs as discussed above. If there exists ϕ^* that satisfies both conditions, the proof of the lemma is straightforward. Since we have $|S|(T+1)$ sets of (s, t) , there will be at least $|S|(T+1)$ non-zero variables due to the condition. Since there is also at most $|S|(T+1) + 1$ non-zero variables in ϕ^* , we can conclude there exists at most one set of (s, t) such that there contains two non-zero variables in ϕ^* .

Below we consider the general case that we can only find ϕ' that satisfies the first condition but not the second. We demonstrate how to construct a “smaller” problem that satisfies both conditions in a smaller problem instance. We then argue the optimal solutions in the smaller problem space is also optimal in the original space. First, we know there always exists a solution ϕ' that satisfies the first condition from the property of linear programs. Now let us construct a “smaller” problem of the original Equation (3.6). For a given ϕ' that satisfies the first condition, denote $Y = \{(s, t) | \forall s, t\}$, the set of all (s, t) , and $X = \{(s, t) | \sum_a \phi'(s, a, t) =$

$0\}$, the set of (s, t) such that $\sum_a \phi'(s, a, t) = 0$. Now construct an updated problem from Equation (3.6) such that the the set of state-time pair is $Y \setminus X$ (i.e., the set of state-time pair that is in Y but not in X), but the transition and cost is still the same, and corresponding action probability is set to zero, i.e., $\phi(s, a, t) = 0$ if $(s', t + 1) \in X$ and $P(s'|s, a) > 0$. Note that ϕ' is still the optimal solution in new problem, and any optimal solution in new problem is also going to be optimal in the original problem (since they perform at least as well as ϕ'). Note that now in the new problem, the number of constraint is $|S|(T + 1) + 1 - |X|$, and the number of state-time pair is $|S|(T + 1) - |X|$. Again, using the property of linear program, there exists a solution with at most $|S|(T + 1) + 1 - |X|$ elements. If we can find a solution ϕ^* that satisfies the above while also satisfying the condition that there exists an action a for every $(s, t) \in Y \setminus X$ such that $\phi^*(s, a, t) > 0$, there exists at most one (s, t) with two nonzero $\phi^*(s, a, t)$ and we have the proof. If not, we can continue the above process to keep shrink the set of (s, t) until we find the new problem that satisfies both conditions. Note since in each new construction, the set of (s, t) is reduced at least by 1, and therefore this procedure will terminate in a finite number of times.

We demonstrates the existence of solution such that "multiple competing nudge" only happens once. Now we show how to leverage the simplex method to find such a solution. The process is essentially an implementation of the procedure in the proof using the simplex method. Given problem 3.6, we can first use the simplex method to find a basic feasible solution, named ϕ^1 . If ϕ^1 satisfies the requirement that "multiple competing nudge" happens at most once, then ϕ^1 is the desired solution. If not, since ϕ^1 is a basic feasible solution, there must exists some (s, t) such that $\phi^1(s, a, t) = 0, \forall a$. We could then reconstruct a problem by removing all those state-time in Equation (3.6), then resolve the problem using simplex method to find a new basic feasible solution ϕ^2 . Note that ϕ^2 has the same performance of ϕ^1 in the original problem. If ϕ^2 meets the requirement, we have found the desired solution. If not, we could repeat above process to construct new problem and find new solution, until the solution satisfies at most one "multiple nudge". Note that each new solution is still an optimal solution in original problem. In the worst case, we need to repeat the above process $|S|(T + 1)$ times, i.e., the number of state-time pair. However, this is still linear in terms of time complexity. By following the above procedure, we can find a solution that "multiple competing nudge" happen at most once in polynomial time.

□

Appendix B

Supplemental Experiment Results

B.1 Additional Experiment Results in Chapter 3

We present two additional sets of simulation results in this section. The first one examines the choice of β in the relaxed formulation in environment design via reward modification. The second one examines the scenario when the agent reward function and biases are not known a priori and evaluate whether we can leverage inverse reinforcement learning to infer agent rewards and biases to be used in our algorithms.

The effect of β in the relaxed formulation. We have shown that solving the environment design problem as defined in Equation 3.3 is NP-hard and have proposed an relaxed formulation as in Equation 3.4. The key parameter of this relaxation is β in the soft-max function. When $\beta \rightarrow \infty$, the relaxation is the same as the original environment design problem. However, in practice, we can only solve it with a finite β .

Here we examine how much the choices of β impacts the outcome. We consider the misalignment of the principal’s and the agent’s reward function. For the agent model, we consider the boundedly-rational agent with $\tau = 1$ and present-bias agent with $k = 1$ (we have examined a range of different parameters, and the results are qualitatively similar.). For comparison, we brute-forcedly derive the optimal solution using bi-level solver of Pyomo¹⁶ and examine how fast the performance of our algorithms converges to the true optimal as β increases. As shown in Figure B.1, the performance converges quickly with β increases. It suggests that setting a small β is enough to reach reasonable approximations. This result also complements Lemma 3, proving the convergence of Q functions, and demonstrates that we can approximate the overall performance of the optimal.

We have also measured the runtime improvements for the relaxation. In our simulation ($|S|=100$, $|A|=4$, and $T=20$), it takes 7.9 seconds for our algorithm to solve an instance on average in our relaxed formulation while it takes 721.3 seconds to solve the instance exactly. The results demonstrate the efficiency improvements of the relaxation.

Unknown agent reward and biases. In our setting, we assume the reward function and agent bias parameters are known. While this assumption might be approximately satisfied in some cases (e.g., reward functions are payments specified by the system, and biases can be roughly estimated as in our experiment), it might not be satisfied in other cases. When

¹⁶<https://github.com/Pyomo/pyomo>

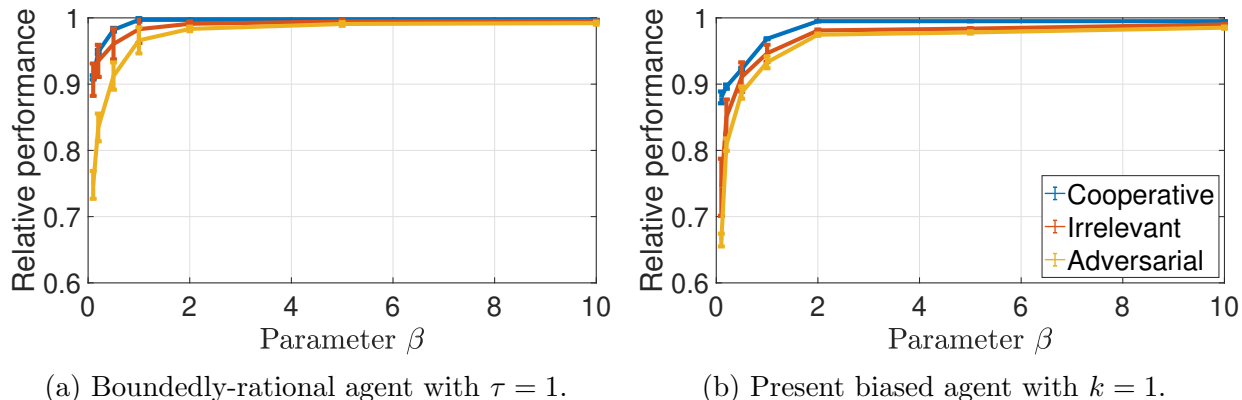


Figure B.1: Examining the impact of β to the relaxed reward function modification algorithm. When β is large, the relaxation is close to optimal. The results suggest that a small β is sufficient for the approximation.

these parameters are unknown, if we have access to data of human behavior in the original environment (e.g., user action history on the website), we might apply standard approaches in inverse reinforcement learning to simultaneously infer the reward and human biases first and then use the inferred values for environment design. In this set of simulations, we examine whether this idea is feasible.

We use the same simulation setup as in the main paper and apply the techniques by [57] to infer the reward and bias parameter at the same time from the policy. Since they take a Bayesian approach, and the prior (initial belief about the parameters) would influence the outcome, we run simulations by assuming the prior is a noisy observation of the truth. In particular, let $r(s, a)$ be the true reward. In the prior, we randomly draw the prior of $r(s, a)$ to be $N(N(r(s, a), \sigma), \sigma)$, where $N(\mu, \sigma)$ is a normal distribution with mean μ and variance σ . Intuitively, larger σ implies a worse prior. We set the agent model to be a bounded rational agent with $\tau = 1$ (we have tried other agent models and the results are qualitatively similar). Figure B.2a and B.2b demonstrate how well the inverse reinforcement learning can estimate the true values with different noise σ in the prior. We then run our environment design algorithms on the inferred values, and Figure B.2c and B.2d demonstrate that our algorithms work on inferred rewards and biases as long as we have reasonable initial prior. While the results in this simulation are exploratory, it showcases the possibility to utilize environment design even when the rewards and human biases are initially unknown.

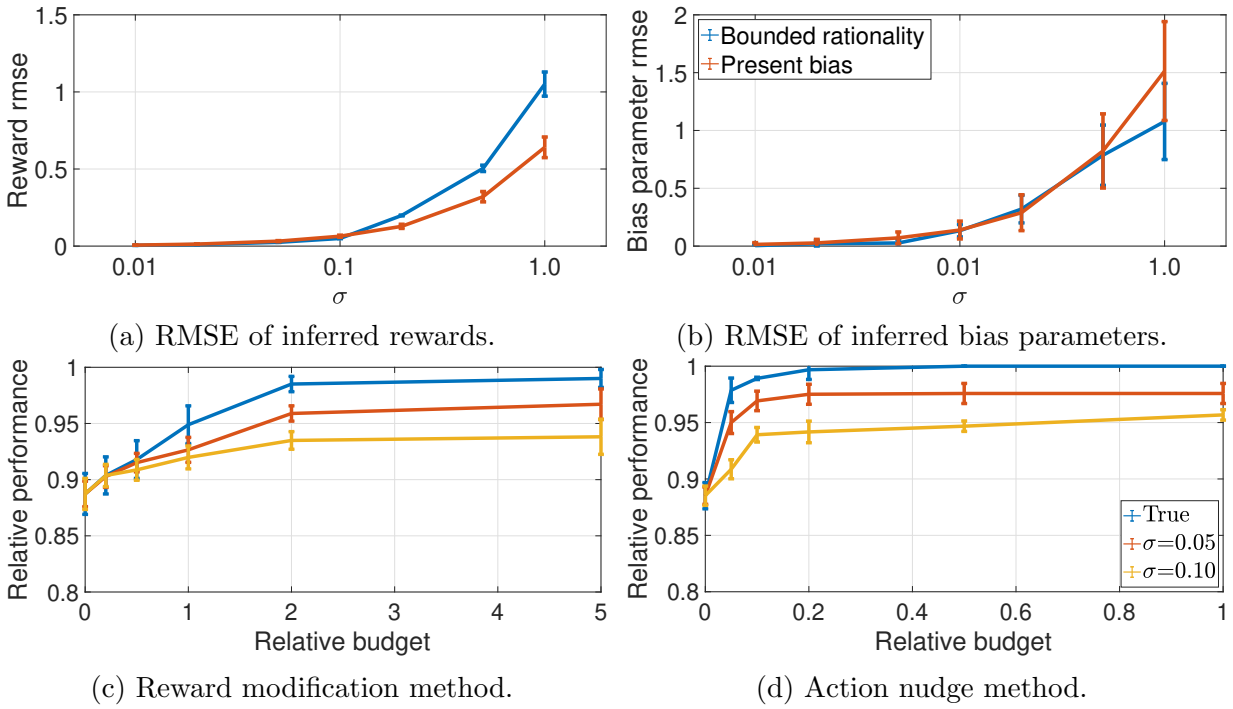


Figure B.2: Performance of reward modification and action nudge methods in environment design problems when the rewards and bias parameters are inferred from observations.

In this chapter, we discuss additional experimental results to evaluate our proposed AI models in the dissertation.

B.2 Additional Experiment Results in Chapter 4

We discuss additional sets of simulation results to highlight the properties and performance of HAIDNet, as well as details of the optimization process of HAIDNet.

B.2.1 Data Generation in Experiments

Here we provide the details in generating data instances for training HAIDNet in our settings.

Single receiver, binary actions and binary states In the simplest setting with binary actions and binary states, the action space is $\mathcal{A} = \{0, 1\}$ and the state space is $\Theta = \{0, 1\}$. We adopt a stylized setting for binary actions where the sender obtains utility 1 when the receiver takes action 1 and utility 0 when the receiver takes action 0 [94]. The receiver utility u^R is uniformly drawn from $[0, 1]$ and prior distribution is draw from Dirichlet distribution. We filter out trivial problem instances where the receiver will always choose one action whatever the information policy, e.g., the receiver always chooses action 1 when receiver utility $u^R(1, \theta) > u^R(0, \theta), \forall \theta \in \Theta$. Total 102,400 instances are generated for training, 1,000 for validation and 1,000 for testing.

Single receiver, multiple actions, and multiple states In the setting with N actions and M states, the action space is $\mathcal{A} = \{0, 1, \dots, N-1\}$ and the state space is $\Theta = \{0, 1, \dots, M-1\}$. The sender utility is set to $u^S(a, \theta) = \frac{a}{N-1}, \forall \theta \in \Theta$ if $N \geq 3$, and the same as above binary actions if $N = 2$. The receiver utility u^R is uniformly drawn from $[0, 1]$ and prior distribution is drawn from Dirichlet distribution. We also filter out trivial cases where the receiver will always choose one action whatever the information policy is.

Multiple receivers, binary actions, and binary states The receiver utility and prior distributions are generated in the same way as in the cases of a single receiver, binary actions, and binary states. The sender utility is the fraction of receivers choosing action 1, i.e., her utility is given $\frac{|S|}{K}$ if there are $|S|$ number of receivers choosing action 1 and K is the total

number of receivers. We also filter out trivial cases where the receiver will always choose one action whatever the information policy is.

Problem instances in human-subject experiments In our human-subject experiment, the problem setup is the same as the setting with a single receiver, binary actions, and binary states. To make the setting easier to understand for experiment participants, the receiver utility is drawn from $\{1, 2, 3, 4, 5\}$ when the participant chooses to purchase a good product or chooses to not purchase a bad product, and the participant utility is 0 for other cases. The sender utility is set to 1 when the receiver chooses to buy, and 0 otherwise. The prior distribution is drawn from the Dirichlet distribution, however, we round all probability in the prior distribution and the information policy to the nearest tenth digit, $\{0\%, 10\%, \dots, 100\%\}$, to make it easier to interpret for human participants.

B.2.2 Convergence and Scalability of Proposed Methods

Convergence of training. In this set of simulations, we have examined the convergence of training with respect to the number of training iterations and also with respect to the softmax parameter β when dealing with Bayesian rational receivers. Overall, HAIDNet converges to finding the optimal policy within reasonable setup.

To illustrate the results, here we present the simplest setting with binary actions and binary states, namely, the action space $\mathcal{A} = \{0, 1\}$ and the state space $\Theta = \{0, 1\}$, and observe whether HAIDNet can produce near-optimal information policies. For the sender utility, we adopt a stylized setting where the sender obtains utility 1 when the receiver takes action 1 and utility 0 when the receiver takes action 0. We randomly draw each value in the receiver utility u^R from $[0, 1]$. The prior distribution λ is drawn from a Dirichlet distribution. We then simulate data using the setting above and optimize HAIDNet.

We compare the performance of the policy learned by HAIDNet with the closed-form optimal policy. Recall that when the receiver is rational (expected utility maximizer), he chooses the action that maximizes his expected utility given his belief about the state. As introduced in Section 4.2.2, to enable the gradient-based method in optimizing HAIDNet, we replace this *argmax* operation as *softmax* using a softmax scale parameter β . Therefore, we first examine the impact of this choice of β and the amount of training (# iterations in gradient descent)

in optimizing the information policy. As shown in Figure B.3, when β is large enough and when we optimize over a large enough number of data batches, the learned information policy from HAIDNet converges to the information policy that achieves near-optimal performance.

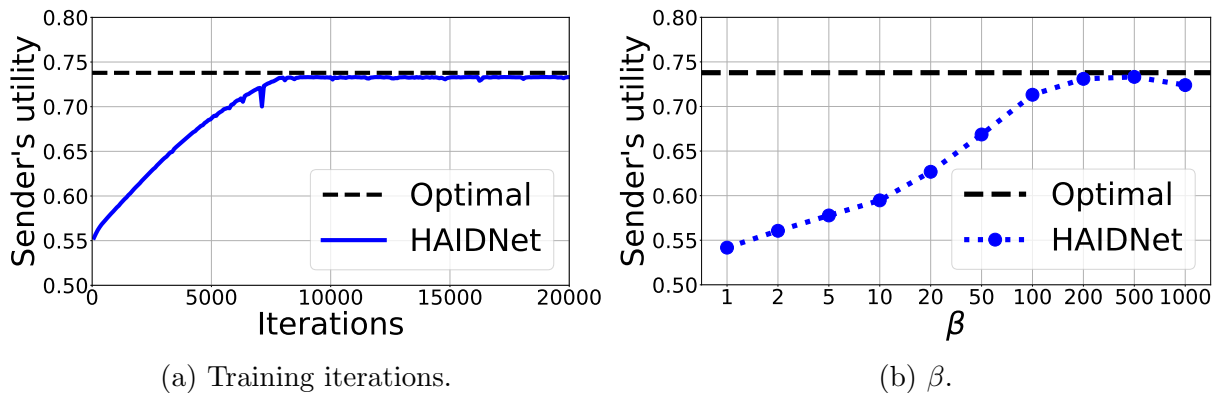


Figure B.3: The convergence results, with respect to the number of training iterations and β , of the sender’s utility derived from the information policy generated by HAIDNet.

Scalability of HAIDNet. One of the benefits of HAIDNet is to provide efficient solutions for settings when it is computational challenging to derive the optimal policy exactly (e.g., in settings with multiple receivers).

To demonstrate this benefit empirically, we first record the time for computing the exact optimal policy for a problem instance with K receivers via a Linear Programming approach [49]. As we can see from Table B.1, the amount of time to compute the optimal information policy grows significantly (the computational complexity grows exponentially as the number of constraints is exponential in the number of receivers in the linear programming approach) as the number of receiver increases. This reaffirms the computational barriers to computing the exact optimal policy. Note that [197] has shown that it is #P-hard to approximate the optimal sender utility within any constant multiplicative factor. So this computational barrier is backed by theoretical analysis.

For HAIDNet, for each class of problems (i.e., a given number of receivers), we only need to train HAIDNet once. For each new problem instance (different priors, sender/receiver utilities, etc), we only need to make a test-time prediction (one pass of forward propagation) to generate an information policy. Again, in Table B.1, we report the time for training HAIDNet and the time for generating the information policy for each problem instance. To provide the number comparisons, when the number of receivers is 18, traditional linear program method of solving the information policy for a problem instance takes more than

23 hours. On the other hand, for HAIDNet, we only need a little more than 1 hour to train HAIDNet for all problem instances with 18 receivers, and it takes less than 1 second to generate the information policy for each receiver. The reported numbers are performed on the machines with Intel(R) Xeon(R) Gold 6148 CPU (2.40GHz) and a Tesla V100-SXM2-32GB GPU.

Table B.1: Comparing run-time between HAIDNet and linear programming methods. K is the number of receivers. The reported run-times are in seconds.

K	Training Time of HAIDNet	Testing time per instance of HAIDNet	Optimal policy per instance via Linear Programming
2	1082	0.184	0.323
3	1291	0.189	0.367
5	1571	0.221	0.371
10	2174	0.270	4.820
15	3284	0.299	235.0
17	3713	0.333	14290
18	4030	0.352	84280

B.2.3 Generalizability of Proposed Methods

Single Bayesian rational receiver case. In Section 4.3, we compare the performance between the policy from HAIDNet and the optimal policy in the single Bayesian rational receiver setting with an increasing number of states with binary actions, and an increasing number of actions with binary states. To further complete the results, we have also run simulations when we simultaneously increase the number of actions and the number of states at the same time. To put the performance of HAIDNet into context, we also include the performance of random policy, which provides random signals all the time. This random policy serves as the naive baseline setting. The results are shown in Table B.2c. The average sender utility obtained by HAIDNet policy is close to optimal policy in both training and testing problem instances (averaged over 1,000 instances) even in cases with large action and state numbers. We also evaluated the model training error for binary action case and binary state case in Table B.2, which shows that HAIDNet works well for large-scale problem instances.

Varying number of receivers, actions and states. In previous sections, we evaluate HAIDNet can accommodate any problem instance (i.e., different specifications of priors,

Table B.2: Comparing the average sender utility generated by the optimal policy and the policy from HAIDNet in the setting with a single Bayesian rational receiver.

(a) Increase the number of states M with binary actions.

M	Training			Testing		
	Random	HAIDNet	Optimal	Random	HAIDNet	Optimal
2	0.4901	0.7409	0.7498	0.4909	0.7408	0.7451
3	0.5009	0.7737	0.7782	0.4819	0.7598	0.7669
5	0.4898	0.8171	0.8209	0.5227	0.8066	0.8225
10	0.4841	0.8495	0.8699	0.4838	0.8196	0.8686

(b) Increase the number of actions N with binary states.

N	Training			Testing		
	Random	HAIDNet	Optimal	Random	HAIDNet	Optimal
2	0.4901	0.7409	0.7498	0.4909	0.7408	0.7451
3	0.4911	0.7017	0.7214	0.5064	0.7089	0.7227
5	0.4919	0.6906	0.7113	0.5119	0.6690	0.7064
10	0.4907	0.6861	0.7084	0.4861	0.6623	0.6963

(c) Increase both the number of states and actions. $M = N$ represents state number equals action number.

$M = N$	Training			Testing		
	Random	HAIDNet	Optimal	Random	HAIDNet	Optimal
2	0.4901	0.7409	0.7498	0.4909	0.7408	0.7451
3	0.4791	0.7352	0.7679	0.4854	0.7199	0.7587
5	0.5029	0.7771	0.8113	0.5101	0.7755	0.8121
10	0.4799	0.8613	0.8971	0.4872	0.8323	0.8994
50	0.4903	0.9247	0.9550	0.7058	0.9166	0.9545

receiver utility, and sender utility) with a fixed number of actions, states, and receivers. It is then natural to wonder whether we can extend the HAIDNet structure so that it can work with varying numbers of receivers, actions, and states. As a proof of concept, in this set of simulations, we attempt to address this question and present an approach that can work with varying numbers of receivers, actions, and states when the numbers are upper bounded.

We first examine the relaxation of a fixed number of receivers. In particular, we can generalize our approach to address varying numbers of receivers when the number of receivers is upper bounded. One straightforward approach is to maintain multiple HAIDNet, one for each fixed number of receivers, for generating the optimal information policy. Another approach is to train a HAIDNet that can generate information policy for the maximum number of receivers. In settings when the number of receivers is less than the maximum number,

Table B.3: Comparing the average sender utility by the optimal policy and the policy from HAIDNet in the setting with at most 10 Bayesian rational receivers.

K	Training			Testing		
	Random	HAIDNet	Optimal	Random	HAIDNet	Optimal
2	0.5042	0.7830	0.8018	0.4986	0.7538	0.7921
3	0.5066	0.7337	0.7586	0.4866	0.7139	0.7450
5	0.5032	0.7245	0.7451	0.5071	0.7121	0.7387
10	0.4944	0.6911	0.7118	0.5009	0.6650	0.6901

we can include “null receivers” who always choose action 0 (by setting the receiver utility such that the utility for taking action 0 is always larger than taking other actions in both states). By including this in the training process, we can have a single HAIDNet that can generate policies for a bounded variable number of receivers. As a proof of concept, we have implemented the above approach and trained a HAIDNet that can work with up to 10 receivers. We then examine its performance when the number of receivers is smaller than 10. As we can see from Table B.3, this approach achieves reasonable performance and shows promising results.

Table B.4: Comparing the average sender utility by the optimal policy and the policy from HAIDNet in the setting with at most 5 states and 5 actions, for a single Bayesian rational receiver.

(M, N)	Training			Testing		
	Random	HAIDNet	Optimal	Random	HAIDNet	Optimal
(2, 3)	0.4994	0.6564	0.7308	0.5276	0.6517	0.7411
(2, 4)	0.4852	0.6450	0.7134	0.5236	0.6535	0.7329
(2, 5)	0.4898	0.6498	0.7042	0.5111	0.6641	0.7258
(3, 2)	0.5094	0.6856	0.7735	0.4731	0.6462	0.7574
(3, 3)	0.5128	0.7072	0.7791	0.4832	0.6689	0.7615
(3, 4)	0.5343	0.7165	0.7729	0.5322	0.6940	0.7672
(3, 5)	0.4798	0.6922	0.7453	0.5308	0.6990	0.7492
(4, 2)	0.4898	0.6849	0.7701	0.5216	0.6922	0.7883
(4, 3)	0.4721	0.7051	0.7761	0.4796	0.6940	0.7844
(4, 4)	0.5032	0.7239	0.7812	0.5143	0.7186	0.7962
(4, 5)	0.4700	0.7347	0.7807	0.5144	0.7421	0.7925
(5, 2)	0.4883	0.7147	0.7915	0.5186	0.7137	0.8038
(5, 3)	0.5394	0.7736	0.8398	0.4928	0.7318	0.8184
(5, 4)	0.4998	0.7810	0.8289	0.4951	0.7494	0.8242
(5, 5)	0.4819	0.7722	0.8159	0.4863	0.7605	0.8079

We now examine whether this approach also works for extending the number of states M and the number of actions N . As a proof of concept, we adopt the same approach above and train a HAIDNet for a maximum of 5 actions and 5 states. We then examine the performance of HAIDNet for problem instances with less or equal to 5 actions or states. As shown in Table B.4, this approach also works in addressing varying numbers of actions and states.

B.3 Additional Experiment Results in Chapter 5

We present additional evaluation of experiment results on human ethical decision-making.

B.3.1 the Effect of Prediction Magnitude

We perform additional exploratory analysis on the collected data in Experiment 1 to gain more insights on how human ethical preferences are affected by predictive information. We look at the impact of not just the direction of the preference in predictions, but the magnitude of prediction differences. Instead of looking at individual factors, we look at how the predictive information impacts human preferences as a whole. More concretely, using the data collected in the first treatment (verifiable only), we can determine the prior preferred candidate, the candidate who is more preferred for each scenario (i.e., a pair of candidates with different combination of factor differences) on the population-level in the first treatment. We then split the scenarios in the second treatment (verifiable and predictive) into 7 groups, where the difference between the survival chance of the prior preferred candidate and the unpreferred candidate is $\{-6, -4, -2, 0, 2, 4, 6\}$. We then measure ΔP of the overall candidate preference (as opposed splitting up by dimension) for each group to understand the impact of the prediction magnitude on the prior preference.

From the results in Figure B.4, we can see how ΔP changes for various magnitudes of prediction value difference. This trend is monotonic, which makes sense intuitively, as we would expect that a larger difference in prediction values has a bigger effect on ethical preferences than a smaller difference in prediction values. However, when the predictions are equal between candidates, workers' ethical preferences decreases compared with the verifiable only group. This result again supports that adding predictive information could impact human ethical preferences, even when the predictive information does not seem to provide differentiating information between candidates.

B.3.2 the Effect of Value Similarity Claims

In our Experiment 3 (in Section 5.3.3), we study the effect of value similarity on AI reliance. To further understand why we see effects of value similarity on AI reliance, we conduct extra

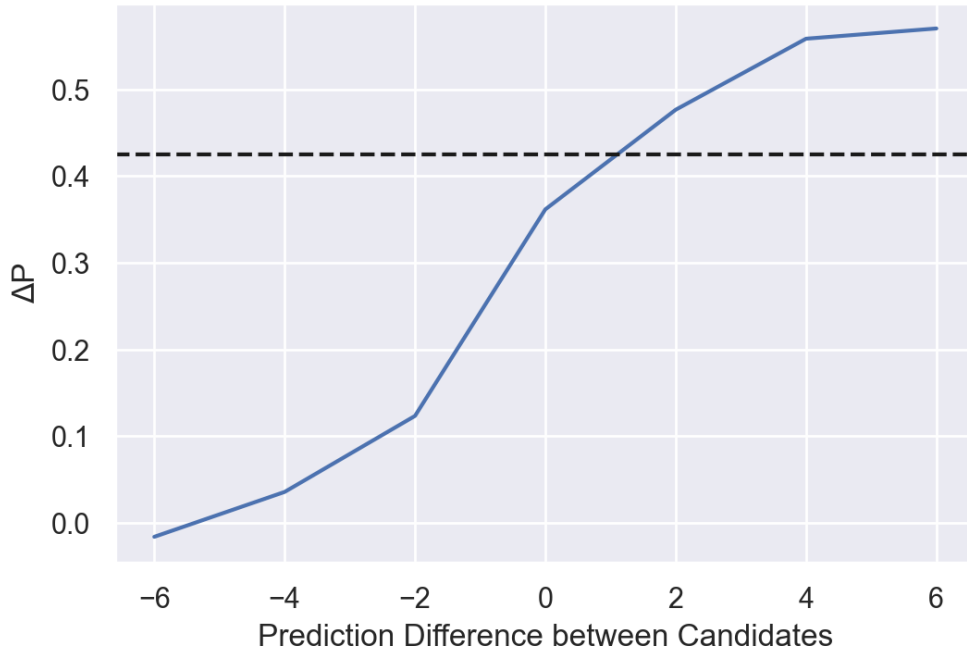


Figure B.4: Effect of prediction magnitude in Experiment 1. We present ΔP for each magnitude of prediction difference in blue, and ΔP for the verifiable only treatment group in black.

data analysis. Specifically, we want to see if the increases in AI alignment caused by value similarity can be explained by the workers’ belief that the AI shares a similar set of values to the workers, or if the increase in AI alignment is due to the actual similarity in values exposed in AI recommendations reinforcing the workers’ own preferences.

In our experiment design, half of the AI recommendations in the second stage are generated deterministically according to the claimed ethical preference, and half of the AI recommendations are generated randomly. When the AI is random, any alignment increase is only due to the perception of the AI having similar or dissimilar values. When the AI is deterministic, alignment increases are explained by both user perception of AI similarity and the effect of the AI actually acting according to its preferences. As a result, we can compare these two to find the isolated effect of AI claims.

We measure the effect of value similarity on conditional AI alignment (as in Section 5.3.3), and break this data down by AI Behavior: whether the AI is deterministic or random. These results are presented in Figure B.5. In this experiment, we have two independent variables (deterministic vs random AI, and similar vs dissimilar AI). The dependent variable

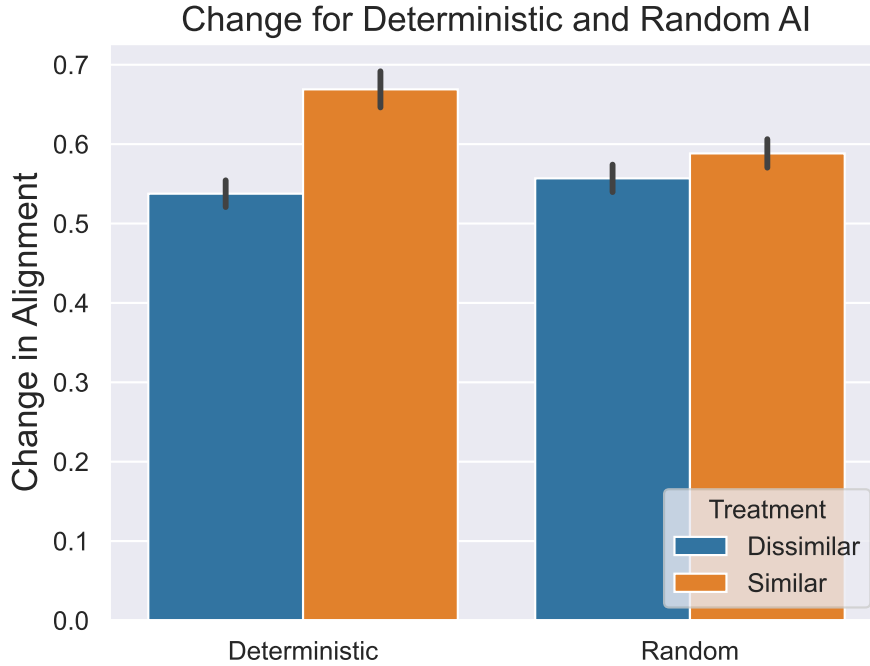


Figure B.5: The effect of value similarity on alignment change between Stages 1 and 2 in Experiment 3, across combinations of Deterministic/Random and Similar/Dissimilar treatments. When the AI is Deterministic, the Similar AI leads to a significantly larger change in conditional alignment ($p < .001$). However, when the AI is Random, there is no significant difference between Similar and Dissimilar AI ($p = .58$).

is the conditional alignment. To examine the significance of the results, we first conduct a two-way ANOVA test and find a significant interaction effect between the two independent variables ($F(1) = 6.86, p = 0.009$). We then conduct post-hoc Tukey’s HSD tests. We find that when the AI is deterministic, there is a significant difference in the conditional AI alignment between similar and dissimilar AI ($p < 0.001$). However, when the AI is random, we see no significance in the conditional AI alignment between similar and dissimilar AI ($p = 0.58$). The results suggest that workers’ reliance on AI is influenced by the realized AI recommendation instead of the value AI claims to exhibit. With this result, we find no evidence to support that the effect of value similarity is primarily due to humans relying on AI recommendations which claim to share similar values, as we see no effect from AI similarity claims alone on reliance.

B.4 Additional Experiment Results in Chapter 6

We conducted additional experiments in a simpler grid world environment with multiple goals to illustrate our study of humans’ beliefs about AI behavior.

B.4.1 Experiment Environments: Grid Worlds with Single Player

We use grid worlds of size of 6×6 in both simulations and human experiments. Similar to the environment setup in Section 6.3.1, the grid world contains a start position, two goal positions, and some blocked positions that the player cannot enter. The player needs to move from the start position towards one of the goal positions. The player can choose to move {Up, Down, Right, Left}. The player will get a positive reward upon reaching the goal, and we set the maximum number of actions to be 20.

We first recruit participants to engage in a single-player game and record their behavior. We then build human behavior model using behavioral cloning. To evaluate our proposed approaches in modeling human beliefs and designing cooperative AI, we conducted additional two sets of experiments. In the second set of experiments, we provide participants with different traces of actions from other agents, again in a single-player game, and ask participants to infer the goal of the agents. This experiment helps us evaluate whether our belief model leads to better predictions of human beliefs over others’ behavior. Afterwards, we conduct a two-player game in the third experiment, where humans are paired with different AI agents to examine the team performance of our design of collaborative AI.

The interfaces of our experiments can be seen in Figure B.6a, B.6b, and B.6c. The detailed descriptions of the experiment and interfaces are included in Appendix C.4.

B.4.2 Experiment 4: Evaluating Human Behavior in Single-Player MDP

Similar to Experiment 1, we recruited 200 workers from Amazon Mechanical Turk. Each recruited worker was asked to play 15 navigation games within a grid world, as shown in Figure B.6a. Our goal is to leverage the collected data to create a data-driven model of human

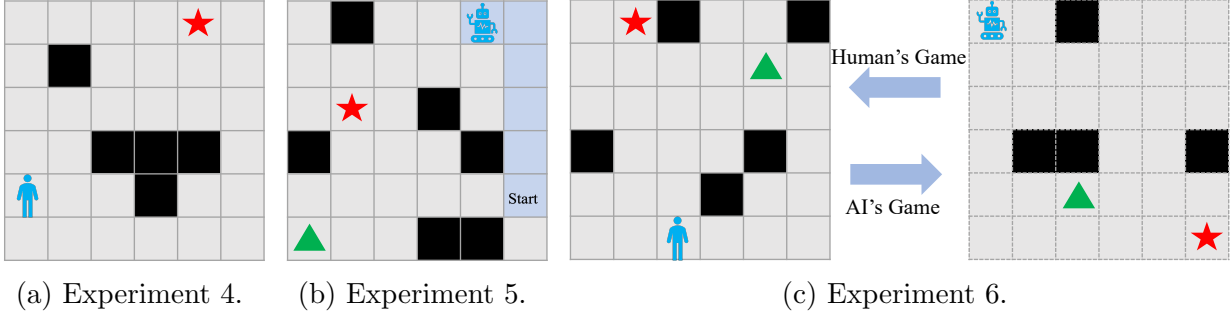


Figure B.6: Human-subject experiment interfaces. In Experiment 4, each participant is asked to control the player to move to the goal (red star). In Experiment 5, each participant is provided a trace of the behavior by another agent, and is asked to infer which goal the agent is trying to reach. In Experiment 6, each participant is playing with an AI agent in separate environments.

behavior. We divided the collected data of human actions into three sets: training, validation, and testing. The training set comprised data from 160 workers, including approximately 70,000 instances of user decisions, while the validation and testing sets each contained data from 20 workers, amounting to around 8,800 instances of user decisions each. The model and parameter tuning were similar to what we did in Experiment 1.

Evaluation of data-driven behavior models. The training, validation, and test accuracies of assuming optimal behavior and behavioral models are presented in Table B.5. Results show that data-driven model could predict human behavior more accurately than assuming humans are optimal.

Table B.5: The prediction accuracy for human behavior for different human models in Experiment 4.

	Training Accuracy	Validation Accuracy	Testing Accuracy
Assuming Optimal Behavior	0.7266	0.6964	0.7131
Data-Driven Model	0.9189	0.8136	0.8422

B.4.3 Experiment 5: Evaluating Human Belief Models

We developed human belief models (behavioral level-1) similar to previous experiments. We examine whether belief inference using behavioral model aligns with real humans, and whether we can design AI behavior such that it is easier for humans to infer the goal of the AI agent.

Experiment setup. The experiment setup is presented in Figure B.6b. The grid world contains a starting position and two goal positions. For each participant, we show them a trace of behavior from another agent and ask the participant to infer which of the two goal the agent is trying to reach.

Experiment 5.1: Examining the belief models. We recruited 200 workers from Amazon Mechanical Turk to compare standard level-1 and behavioral level-1 model. Each worker was asked to infer the goal for 25 behavioral traces. We compared the performance of two belief models and the worker accuracy, as shown in Table B.6. As we can see from the table, humans are generally poor at inferring the goal of other agents: their accuracy only reaches 59.37% in inferring the goal of the other agent. Moreover, the behavioral level-1 model, which accounts for the human behavior model in the belief model, captures human beliefs better than the standard level-1 model, which assumes human behavior is optimal.

Table B.6: Performance comparison between Bayesian inference framework using standard model and human behavior model.

	Consistency to Human Predictions	Cross Entropy Loss
Standard level-1	0.4977	0.9284
Behavioral level-1	0.5764	0.7506
True goals	0.5937	0.6631

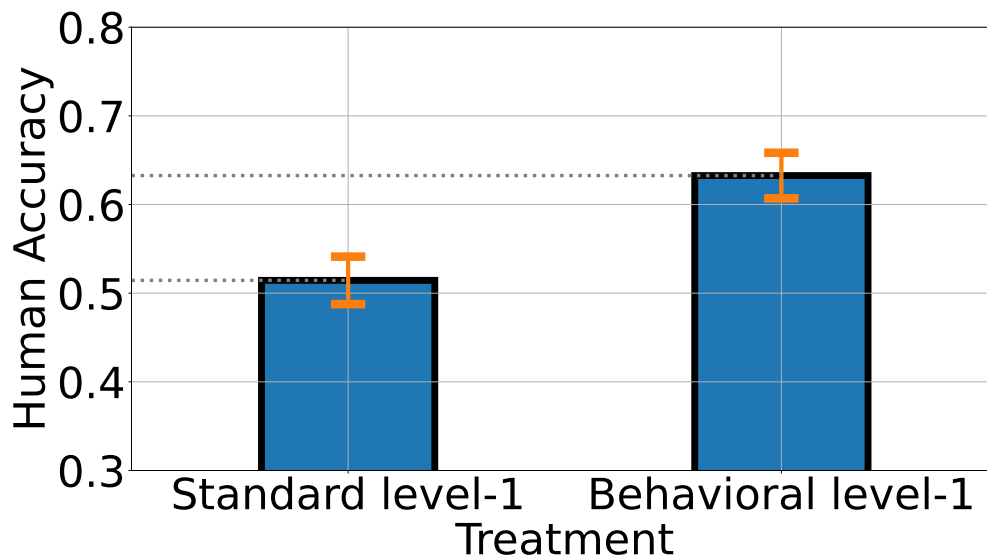


Figure B.7: Human evaluations of belief inference accuracy regarding AI goals.

Experiment 5.2: Developing explicable AI policy. As demonstrated in Experiment 5.1, humans generally struggle to infer the goals of other agents based on others' behavior.

To design AI agents which makes inference of their goals easier by humans, we train AI agents to maximize the likelihood that humans can accurately infer these goals from their behavior.

We recruited 400 workers from Amazon Mechanical Turk, randomly assigning them to one of two treatments, to assess the behavior of AI agents. Each participant was tasked with identifying the goals of the player in 30 different scenarios. Participants were awarded a \$0.03 bonus for each correctly identified goal. Figure B.7 displays the results of human evaluation. The findings indicate that humans achieve higher accuracy in inferring the correct goals of AI agents when employing Behavioral level-1 model.

B.4.4 Experiment 6: Evaluating Collaborative AI Agents

Utilizing developed models of human behavior and beliefs, we follow the same methodology in Experiment 3: train different collaborative AI agents, pair them with different human models, and examine the collaborative performance in simulations and human subject experiments (via recruiting 300 workers).

The setup of our Experiment 6 is shown in Figure B.6c. The goal for the human-AI team is for both agents to reach the same goal in their own environments (both reaching “red star” or both reaching “green triangle”) within a time limit. The team will not get points if they reach different goals or one of the players fails to reach any goal.

Simulations results are shown in Table B.7. Simulations indicate that collaborative performance is highest when an AI agent is paired with the human model used to train it. Figure B.8 presents the average collaborative reward in human-subject experiments. Statistical analysis revealed significant differences in the performance of AI models when paired with human participants, with p -values of 0.0166 for comparisons between the treatments *self-play* and *Behavior-AI*, and $p < 0.0001$ for *Behavior-AI* versus *Belief-AI*. These results demonstrate that incorporating human beliefs into the design of AI agents enhances collaborative performance when working with real humans.

Table B.7: Simulation results of human-AI collaborative rewards in Experiment 6. Columns players are different AI agents, and row players are different simulated human models.

Human Model	AI Agent		
	Self-play	Behavior-AI	Behavior&Belief-AI
Self-play AI	0.7828	0.4780	0.6164
Behavior	0.5926	0.7584	0.4552
Behavior&Belief	0.6919	0.7268	0.7813

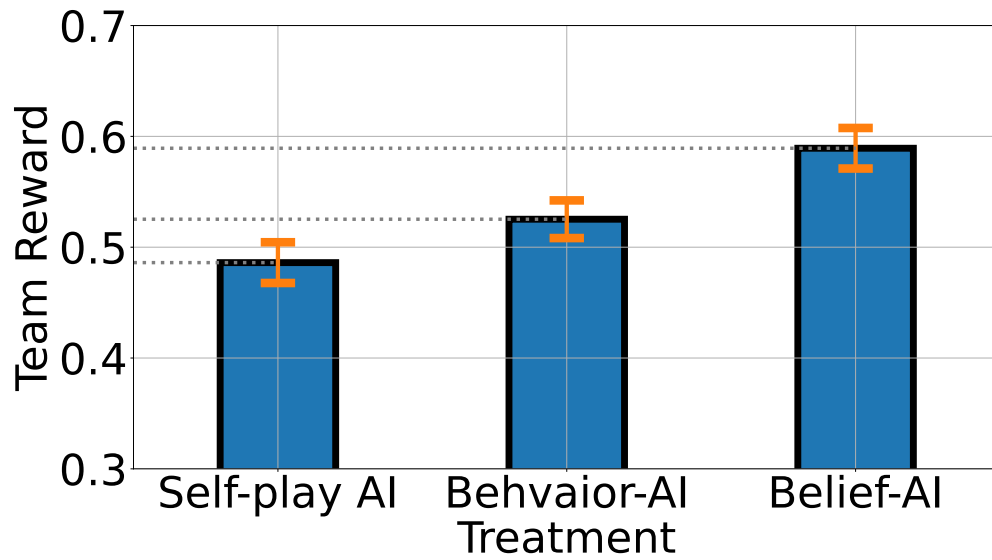


Figure B.8: Average collaborative reward of humans and AI agents in Experiment 6.

Appendix C

Details of Human-Subject Experiments

We provide detailed information about our human-subject experiments in this chapter.

C.1 Human Experiment Details in Chapter 3

In Chapter 3, we study the problem of environment design, and we provide more detailed description of our human subject experiments in this section.

Experiment design. Our human-subject experiment is approved by the IRB board in our institution. In our human-subject study, each worker is asked to play six navigation games, representing the decision-making environments. Similar to our simulation setting, each navigation game is represented by a grid world of size 10×10 . The initial state is in the middle of the grid world, and the time horizon T is set to 20. In order to reduce the cognitive burden for human subjects, the reward function is simplified to only depend on the state, and we let the principal’s reward function to be equal to the agent’s reward function, i.e., $R^a(s, a) = R^p(s, a) = R(s)$.

The reward on each state is an integer from 1 to 100. Similar to the setup in the simulation, we place a high reward state (uniformly drawn from 80 to 100) in a random corner of grid as global optimal, and a medium reward state (from 50 to 80) in other three corners as local optimal. Since the initial point is in the middle of map, we set the reward of path to global optimal and one local optimal to be low (from 10 to 30), and reward towards other two local optimal to be relatively high (from 30 to 50). Other places of map is set to be relatively low (from 1 to 30).

Experiment interface. The interface of the navigation game is shown in Figure C.1. Workers can move the plane around the map to collect the points in the grid world, and their bonuses depend on the total points they collected for the six games. For every 100 points collected, they can earn an additional USD \$0.01 bonus. Taking into account the workers’ working time in the task, and the \$0.50 base payment for submitting the task, the average hourly rate is around USD \$11.50.

To induce biased human behavior, at each time step, a worker can only see the rewards of the nearby states (to simulate the short-sightedness). Out of six games, there are two games each for vision length of 1, 2, 3, which we use short-sighted agent with $\tau = 0, 1, 2$ to model

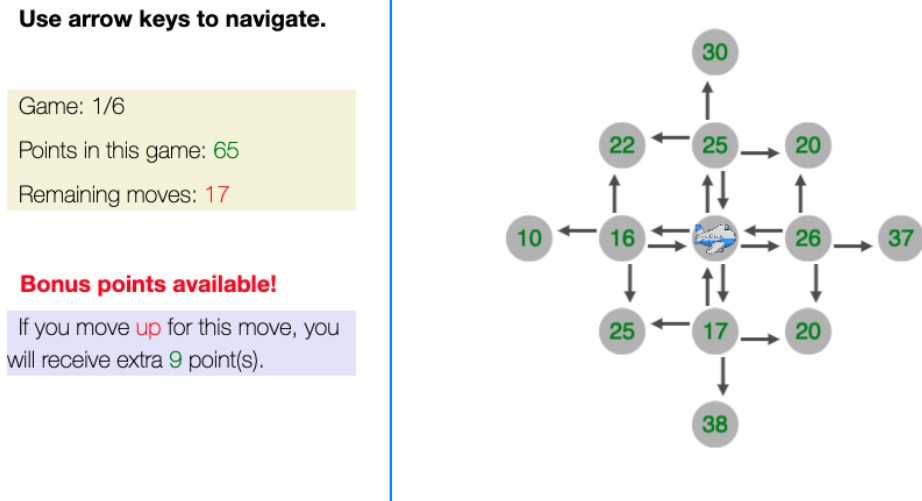


Figure C.1: Human experiment interface of updating decision-making environments in Chapter 3. Workers can use arrow keys to move the airplane around and collect points. The information on the bottom of the left-panel is the action nudge presented to workers, which is shown in action nudge treatment when a nudge is provided by AI model, hidden otherwise.

when solving the environment design problem. Note that the purpose of this design is to provide us an estimate of human biases to be used in environment design. Worker behavior might not follow the behavior model.

Each worker is randomly assigned to one of the three treatments: {baseline, modified reward, action nudged}, with 106, 86, 108 workers assigned to each. The games are drawn from the same pool for each treatment. In the baseline treatment, workers will play the drawn games without modifications. In the modified reward treatment, workers will see the modified rewards generated by our algorithm, while in the action nudge treatment, when the nudge happens, the workers will see an additional message indicating they might gain bonus for moving towards a certain direction (as shown in Figure C.1).

C.2 Human Experiment Details in Chapter 4

In Chapter 4, we study the problem for information design and run human-subject experiments to evaluate designed policy. We compare average sender utility of different policies

in human-subject experiment in Figure 4.4, and we also compute the receiver utility in each treatment, included in Table C.1. As we can see from the table, while HAIDNet helps find a policy that leads to the highest sender utility, it comes at the cost of reducing the receiver utility, a demonstration of the ethical concerns as discussed in Section 4.4.

In our experiment setup, given the sender’s goal is to have the receiver purchase the products regardless of the product quality, when the sender is more successful, it leads to a lower receiver utility in general and implies the potential negative social impacts.

Table C.1: Comparing sender and receiver utility of different policies in human-subject experiments of designing information policy.

Information Policy	Random	BR-Policy	TH-Policy	HAIDNet
Sender Utility	0.489	0.524	0.621	0.667
Receiver Utility	0.663	0.634	0.565	0.532

Experiment interface and description. In our human-subject experiments, we simulate the setting with binary actions and binary states. In particular, we present the product purchasing example as we discussed in Section 4.2.1. The task interface about our human-subject experiments is shown in Figure C.2.

Each human participant is asked to make multiple rounds of purchase decisions. In each round, the participant is presented a product with unknown binary quality (either good product or bad product). The participant is told that a (noisy) inspection has been performed on the product, and is given the conditional distribution associated with the inspection (i.e., the probability to receive a good/bad signal given the product is good/bad). Finally, the participant is given a realization of the inspection signal and is asked to make a binary decision of purchasing or not. The participant’s reward depends on both their purchasing decisions and the true product quality. When collecting human response in the first phase, random policy are presented to all participants. In the second phase, different policies are presented: {Random, BR-policy, TH-policy, HAIDNet policy}. The policies are designed with the assumption that the sender is persuading human receivers to purchase the product, and we calculate the probability of participants choosing to purchase and report it as the sender utility to evaluate performance of different policies.

In this round, consider the following scenario:

1. You are presented with a product . The product could be Good or Bad , and you need to decide whether to purchase it. Your bonus will depend on your purchase decision and the product quality as shown below:

- ★ If you "Buy" a Good product, you will get a bonus of 3 cents.
- ★ If you "Don't Buy" a Bad product, you will get a bonus of 2 cents.
- ★ Otherwise, you won't get any bonus for this round.

	Buy	Don't Buy
	3	0
	0	2

Bonus table

2. Below is the information about the product for you to make the decision.

- ★ The product is randomly selected from a pool of products. For all products, 70% are Good and 30% are Bad.
- ★ After given the product, you will perform a noisy inspection.
 - If the product is good, you will receive a good signal with 20% chance or a bad signal with 80% chance.
 - If the product is bad, you will receive a good signal with 80% chance or a bad signal with 20% chance.

3. Which action will you take when the inspection result is ?

Your Decision Is: Buy Don't Buy

Figure C.2: Human experiment interface of designing information policy in Chapter 4.

C.3 Human Experiment Details in Chapter 5


In Chapter 5, we study the problem of presenting predictive information and recommendations for human ethical decision-making. Besides experiment interface in Figure 5.1, the other experiment interfaces are shown in Figure C.3.

C.4 Human Experiment Details in Chapter 6

In Chapter 6, we study the problem of modeling human belief and conduct multiple real human experiments. The experiment interface is shown in Figure C.4.

Question 1 of 29

Which candidate should receive the kidney transplant first?

	Patient A	Patient B
Kidney Donor Status	Prior Kidney Donor	Not Prior Donor
Wait Time	4 years	5 years
Kidney Disease Stage	Stage 5 (Kidney failure or near-failure)	Stage 4 (Severe kidney damage)
 AI Prediction of Survival Chance	79% chance of survival after 5 years post-transplant	81% chance of survival after 5 years post-transplant

Select: Patient A


Select: Patient B

Please make your selections.
Click the buttons or use the ←/→ keys.

(a) Experiment 2 interface: humans make decisions with predictive information from AI systems or human experts .

Which candidate should receive the kidney transplant first?

	Patient A	Patient B
Kidney Donor Status	Not prior donor	Prior kidney donor
Wait Time	3 years	2 years
Disease Stage	Stage 5 (Kidney failure or near-failure)	Stage 4 (Severe kidney damage)


AI Suggestion

✓

Select: Patient A

Select: Patient B

Please make your selection.
Click on the buttons or use the ←/→ keys.

(b) Experiment 3 interface: humans make decisions with suggestions from different AI systems.

Figure C.3: Human experiment interface of human ethical decision-making in Chapter 5.

Instructions

Games : 7/33

This game is a formal game. Your performance will be taken into account for your bonus. You will get bonus if you and your teammate reach the same goal at the same time or within 2 actions.

Your current position is marked as **■**, and your teammate is marked as **□**.

Your goals are marked as **★/▲**.

Black grids are Walls that the player cannot pass.

Use keyboard to move, or press 'space' to stay.



Instructions

Games : 6/33

This game is formal. You will get bonus if your answer is correct.

The position of two players are marked as **■** and **□**.

Click the button below to replay player's actions.

Replay the actions

Which goal do you believe the **■** player is going to? *

Goal **★**

Goal **▲**

Next game



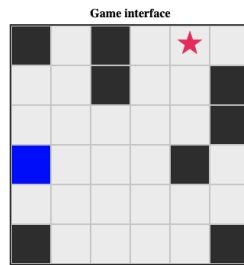
(a) Experiment 1 & 3 interface: humans play games with AI agents in grid worlds of size 8×8 . (b) Experiment 2 interface: humans infer goals of other players in grid worlds of size 8×8 .

Game Information

Your current position is marked as **blue**, and your goal is the **★** grid. You cannot enter Black grids.

Use keyboard to move.

Games : 1/15



Information

Games : 1/30

The current position of the player is marked as **blue**, the trajectory is marked as **light blue**, and goals are marked **red / green**. Black grids are Walls that the player cannot pass.

Click the button below to replay player's actions.

Replay the actions

Which goal do you believe the player is going to? *

Goal **★**

Goal **▲**

Next game



(c) Experiment 4 interface: humans play navigation games in grid worlds of size 6×6 . (d) Experiment 5 interface: humans infer goals of other players in grid worlds of size 6×6 .

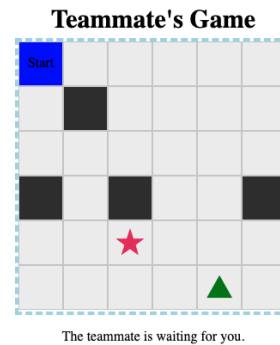
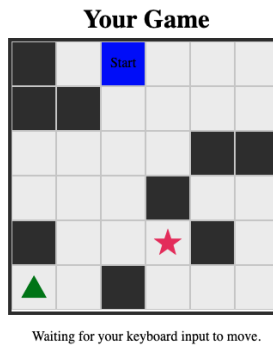
Instructions

Games : 4/23

This game is a formal game. Your performance will be taken into account for your bonus. You will get bonus if you and your teammate reach the same goal at the same time or within 2 actions.

Your current position is marked as **blue**, and goals are marked **red / green**. Black grids are Walls that the player cannot pass.

Use keyboard to move, or press 'space' to stay.



(e) Experiment 6 interface: humans play games with AI agents in grid worlds of size 6×6 .

Figure C.4: Human experiment interfaces of goal navigation and belief inference tasks in Chapter 6.

C.5 Demographic Information of Human-Subject Experiments

For completeness, we include the number of workers in each experiment and their demographic information below. We recruited workers from Amazon Mechanical Turk for our human-subject experiments. Table C.2 lists the number of workers, the number of tasks per worker and the number of treatments in each experiment, and Table C.3 contains the demographic information of all the workers.

Table C.2: Task setup of each human-subject experiment.

Chapter	Experiment	Workers	Tasks	Treatments	Base Payment	Bonus per Task
Chapter 3	Exp. 1	300	6	3	\$0.50	\$0.01 per 100 points
Chapter 4	Exp. 1	100	20	1	\$0.50	\$0.05
	Exp. 2	600	20	4	\$0.50	\$0.05
Chapter 5	Exp. 1	600	29	2	\$0.80	\$0.00
	Exp. 2	300	29	2	\$0.80	\$0.00
	Exp. 3	300	18	2	\$1.00	\$0.00
Chapter 6	Exp. 1	190	30	3	\$1.00	\$0.05
	Exp. 2	200	30	2	\$1.00	\$0.03
	Exp. 3	200	30	3	\$1.00	\$0.05
	Exp. 4	200	15	1	\$1.00	0
	Exp. 5.1	200	25	1	\$1.00	0
	Exp. 5.2	400	30	2	\$1.00	\$0.03
	Exp. 6	300	20	3	\$1.50	\$0.05

Table C.3: Demographic information of all the participants in our human-subject experiments.

Group	Category	Chapter 3	Chapter 4	Chapter 5	Chapter 6
Age	20 to 29	112	88	465	775
	30 to 39	110	111	340	642
	40 to 49	35	65	243	146
	50 or older	43	36	152	127
Gender	Female	121	131	435	547
	Male	168	161	733	1118
	Other	5	1	32	25
Race / Ethnicity	Caucasian	196	240	975	1511
	Black or African-American	18	18	149	82
	American Indian/Alaskan Native	7	5	35	32
	Asian or Asian-American	63	22	16	28
	Spanish/Hispanic	7	6	13	7
	Other	9	9	12	30
Education	High school degree	20	12	67	60
	Some college credit, no degree	19	9	32	42
	Associate's degree	13	24	37	39
	Bachelor's degree	216	223	817	1217
	Graduate's degree	29	29	230	293
	Other	3	3	17	39
Total		300	300	1200	1690