Research Statement: Behavior-Informed Machine Learning

Chien-Ju Ho

December 20, 2023

Machine learning (ML) has integrated into various facets of human everyday life, largely deriving its training from human data. Consequently, these ML systems often exhibit and reflect human behavioral biases, leading to a host of concerns in applications ranging from social media to medical decision-making. While these concerns underscore the pressing need to factor in human behavior when developing ML systems, current ML methodologies mostly either view humans as independent, stochastic data sources or assume that humans are rational decision-makers, despite substantial evidence from psychological studies indicating that human behavior frequently deviates from these models. Such discrepancies highlight the existing gap in incorporating empirically-grounded human behavior insights from psychology into the design of ML systems. Furthermore, as the capacity of ML and our understanding of human behavior continue to grow, it opens up the rich potential of designing ML systems that augment human decision-making, especially in high-stakes or ethically-sensitive domains where humans are still desired to be the final decision-makers.

My research aims to develop behavior-informed machine learning, examining and incorporating empirically-grounded human behavior into the design of ML systems. I focus on two key aspects of human behavior in the ML lifecycle: the generation of data used for training ML models, and human decision-making in tandem with machine assistance. Correspondingly, my research addresses two principal forms of interactions between humans and ML systems: designing ML systems that learn from human data, and designing ML systems that assist humans in decision-making.

Behavior-Aware ML: Learning from Humans

One major line of my research has focused on how to acquire data from humans for developing machine learning systems. Traditional approaches often make strong assumptions about human behavior during data generation, e.g., assuming humans are stochastic data sources, and assume the dataset is fixed and given. My research aims to relax these assumptions by incorporating empirically grounded human behavior, improving the data collection process, and designing behavior-informed learning algorithms. Below, I highlight a few of my research projects along these directions.

Understanding human behavior through behavioral experiments. Most of the work on the study of systems with humans in the loop assumes simple human behavior models that often fail to represent human behavior in practice. To incorporate empirically grounded human behavior into ML, I have conducted a range of human-subject experiments to examine and understand human behavior during the data generation process. For example, with Alex Slivkins, Sid Suri, and Jenn Wortman Vaughan [7], I examined how online workers react to different performance-based payments. By conducting a comprehensive set of experiments on Amazon Mechanical Turk with more than 2,000 workers, I developed a worker behavior model which introduces the concept of *workers' priors* into the standard economic model. Workers' priors describe workers' beliefs on their probability of getting different payments given their performance. I showed that this model is consistent with our results and the results of previous studies. In addition to financial incentives, I have also empirically examined human behavior in different task designs [15] and when workers are communicating with each other [27, 14], relaxing the standard data independence assumption.

Another important aspect of human behavior during data collection is humans' awareness of the existence of ML. As ML becomes more ubiquitous, human behavior might evolve accordingly. For instance, if users are aware that their movie ratings will impact future recommendations, they may alter their rating behavior. Together with Lauren Treiman and Wouter Kool [28], I examined whether human behavior changes when they are aware that their actions will be used to train ML systems. Using the classical ultimatum game as the decision-making task, we found that humans are more willing to sacrifice their own personal gains to improve the fairness of downstream ML systems when they are aware of the ML training. Moreover, this behavior change is robust, persisting regardless of whether the humans will interact with the trained ML in the future.

Improving data collection: Towards data-centric ML. Data has become the driving force behind the rapid progress of ML. While numerous efforts have been made to advance ML by developing sophisticated models and algorithms assuming the data is fixed, much less attention has been given to intervening in the data collection processes to improve data quality from the outset. One line of research has contributed to data-centric ML, focusing on improving the data used to train ML systems. In particular, my earlier works with Shahin Jabbari and Jenn Wortman Vaughan [9, 6] have explored the problem of assigning heterogeneous labeling tasks to workers and optimally aggregating the obtained labels. Leveraging online primal-dual techniques, I have developed algorithms that learn workers' skill levels through historical records, assign tasks to workers with suitable skills, and smartly aggregate labels based on what we learned. These algorithms are theoretically shown to achieve near-optimal performance and have been empirically demonstrated to perform well with real-world crowdsourcing workers. Notably, the online primal-dual techniques developed are general techniques. I have later applied them to other societal resource allocation problems such as kidney allocation [17] and homelessness prevention [13].

I have also studied the design of incentives to motivate high-quality data collection from humans. With Alex Slivkins and Jenn Wortman Vaughan [8], I explored the problem of learning optimal performance-based payments in crowdsourcing markets, where workers' payments depend on the quality of their work. I extended the standard principal-agent model from economic theory to a multi-round online model and designed a novel bandit algorithm which only observes limited information from workers but can perform nearly as well as an oracle algorithm that has access to full information. In addition to financial incentives, I have explored the design of other forms of incentives, such as reputation systems [11, 16], attention [18], and social verification [4] and implemented human computation games for collecting data from real-world users [3, 2, 10].

Learning from behavior data. In addition to understanding human behavior and improving data collection, I have developed learning algorithms that explicitly account for human behavior when learning from human data. My earlier works have focused on the case of strategic human behavior. With Jacob Abernethy, Yiling Chen, and Bo Waggoner [1], I explored the problem of actively purchasing data from users for solving machine learning tasks. Users are only willing to share their data if the offered prices are higher than their private costs. I showed how to convert a large class of machine learning algorithms into online posted-price mechanisms and learning mechanisms. These proposed mechanisms identify the 'importance' of each data point and decide the payment to offer each user. I proved that our mechanisms are 'incentive-compatible,' meaning that workers are willing to truthfully report their costs. Furthermore, I showed that our mechanisms incur significantly lower costs while achieving learning accuracies of the same order as purchasing all data points. With Rafael Frongillo and Yiling Chen [5], I explored the problem of eliciting workers' confidences to achieve optimal label aggregation, with an additional focus on designing multiple-choice questions. I developed a Bayesian framework to model the process of eliciting and aggregating data from the crowd, providing an incentive-compatible payment scheme, a principled way of aggregating labels, and optimal designs for multiple-choice questions.

I have also incorporated psychology-grounded human behavior into machine learning. With Wei Tang [21], I addressed the problem of bandit learning with biased human feedback, a form of reinforcement learning with human feedback (RLHF). In particular, I considered the setting in which, when eliciting feedback from humans, their feedback is not independently drawn as often assumed in bandit learning. Instead, it is influenced by other users' feedback, known as herding behavior. By formally incorporating this human behavior into the bandit learning framework, we theoretically demonstrate that under certain mild conditions, learning might become infeasible even with an infinite amount of data. This observation reinforces the need for my research in both better understanding human behavior and improving data collection from the start. In addition to incorporating specific human behavioral models, I have also demonstrated the use of robust optimization techniques to design decision rules that remain robust in situations where human models are unknown a priori. This approach is applicable to a general set of human models [26].

Behavior-Aware ML: Assisting Humans in Decision Making

Humans often make suboptimal decisions and engage in 'on-the-job training,' i.e., learn to make better decisions while making them. Conversely, the rapid advancements in ML highlight its potential to enhance human performance and expedite their learning with ML assistance. Another line of my research efforts has been on investigating approaches to understand human decision-making with ML assistance, designing ML assistance to improve human decision outcomes, and examining the downstream impacts of machine learning.

Understanding human responses to ML assistance. In order to design assistive ML systems, we need to gain insights into how humans respond to ML assistance. One framework to address human response to ML assistance is information design [12], where humans incorporate ML assistance as new information to update their beliefs about the world and then make decisions based on these updated beliefs. However, in the vast majority of the information design literature, humans are often assumed to be Bayesian, incorporating information and updating beliefs in a Bayesian manner, and rational, taking actions that maximize their expected utility. With Wei Tang [22], I developed an alternative framework for information design based on the discrete choice model and probability weighting. I conducted online behavioral experiments on Amazon Mechanical Turk and demonstrated that our framework better explains real-world user behavior. With this framework, in my later works, I also investigated the theoretical characterization and optimization methods for the optimal policy.

I have also examined factors that impact humans' reliance on ML assistance, specifically when humans decide to follow recommendations made by ML algorithms. With Saumik Narayanan, Guanghui Yu, Wei Tang, and Ming Yin [20, 19], we conducted a series of human-subject experiments in the context of ethical decision-making, using kidney allocation as an example. We found that the presence of predictive information significantly changes how humans consider other information and that the source of the predictive information (e.g., whether the predictions are made by ML or humans) plays a crucial role in how humans incorporate this predictive information. Moreover, when humans and ML recommendations disagree, humans are more likely to change their opinion if the ML displays similar 'ethical values' as human decision-makers. These projects enhance our understanding of how humans respond to ML assistance, which in turn aids in designing better ML assistance policies.

Designing assistitve ML. My recent work has also begun to address the question of designing ML that assists human decision-making. With Guanghui Yu [29], I investigated the setting in which a (potentially biased) human decision-maker operates in a sequential decision-making environment. Our goal is to design ways to either update the decision-making environment or provide recommendations in an online manner to improve the overall decision outcome. We formulated this problem under the Markov decision process (MDP) framework and incorporated common models of biased agents by introducing general time-discounting functions. We then formalized the environment design problem as constrained optimization problems and proposed corresponding algorithms. Our methods have been shown to be effective in both simulations and real human-subject experiments with workers recruited from Amazon Mechanical Turk.

I also investigated the design of ML assistance through the framework of information design, i.e., how to provide information that leads to the desired outcome. As highlighted earlier, the vast majority of literature in information design often makes strong assumptions about human behavior, and I have proposed an alternative framework with empirically-grounded human behavioral models [22]. With Yiding Feng and Wei Tang, I theoretically characterized the (approximately-)optimal information policy within this framework. Moreover, we also proposed a 'rationality-robust' information policy, where the provided information performs well even when we do not have full information about human behavior. While our results have extended information design to settings beyond the standard assumption of human rationality, they still only address a subset of alternative human models. With Guanghui Yu, Wei Tang, and Saumik Narayanan [30], I developed a data-driven optimization framework that can work with any provided human models, including those where we do not have a closed-form expression of human behavior but have access to human behavioral data. Through extensive simulation, we demonstrated that our data-driven optimization approach not only recovers near-optimal information policies with known analytical solutions but also extends to designing information policies for computationally challenging settings or those without known solutions in general. Through human-subject experiments, we also demonstrated that our approach can capture human behavior from data and lead to more effective information policies for real-world human decision-makers.

Ethical considerations. I have also investigated various ethical considerations related to deploying machine learning algorithms in societal domains. As one prominent example, with Wei Tang and Yang Liu [25], I examined the long-term impacts of actions in sequential decision-making. In the context of loan approval, a bank should not only consider the predicted payback rate of applicants from a disadvantaged group but also assess whether approval decisions can help improve the group's social status in the long term. It is important to note that this consideration isn't solely for promoting fairness; considering the long-term impact of actions could also increase the payback rate from people in the group, ultimately enhancing the bank's long-term utility. This project formalized the concept of long-term impacts of actions in bandit learning and explored algorithmic designs to understand the tradeoff between maximizing immediate payoffs and long-term impacts. In addition to this project, I have addressed other aspects of ethical considerations in the deployment of machine learning algorithms, including ensuring the privacy of various stakeholders when learning algorithms rely on human-generated data [23, 24].

References

[1] Jacob Abernethy, Yiling Chen, Chien-Ju Ho, and Bo Waggoner. Cost-efficient learning via active data procurement. In *ACM Conference on Economics and Computation (EC)*, 2015.

[2] Chien-Ju Ho, Tao-Hsuan Chang, Jong-Chuan Lee, Jane Yung jen Hsu, and Kuan-Ta Chen. Kisskissban: A competitive human computation game for image annotation. In *Human Computation Workshop (HCOMP)*, 2009.

[3] Chien-Ju Ho, Tsung-Hsiang Chang, and Jane Yung jen Hsu. Photoslap: A multi-player online game for semantic annotation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2007.

[4] Chien-Ju Ho and Kuan-Ta Chen. On formal models for social verification. In *Human Computation Workshop (HCOMP)*, 2009.

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

[6] Chien-Ju Ho, Shahin Jabbari, and Jennifer Wortman Vaughan. Adaptive task assignment for crowdsourced classification. In *International Conference on Machine Learning (ICML)*, 2013.

[7] Chien-Ju Ho, Aleksandrs Slivins, Siddharth Suri, and Jennifer Wortman Vaughan. Incentivizing high quality crowdwork. In *International World Wide Web Conference (WWW)*, 2015. **Nominee for Best Paper Award**.

[8] Chien-Ju Ho, Aleksandrs Slivkins, and Jennifer Wortman Vaughan. Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems. In *ACM Conference on Economics and Computation (EC)*, 2014.

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

[10] Chien-Ju Ho, Chen-Chi Wu, Kuan-Ta Chen, and Chin-Luang Lei. Deviltyper: A game for captcha usability evaluation. In *ACM Computers in Entertainment*, 2011.

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

[12] Bolin Ding, Yiding Feng, Chien-Ju Ho, Wei Tang, and Haifeng Xu. Competitive information design for pandora's box. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2023.

[13] Zehao Dong, Sanmay Das, Patrick Fowler, and Chien-Ju Ho. Efficient nonmyopic online allocation of scarce reusable resources. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2021.

[14] Xiaoni Duan, Chien-Ju Ho, and Ming Yin. Does exposure to diverse perspectives mitigate biases in crowdwork? an explorative study. In *AAAI conference on human computation and crowdsourcing (HCOMP)*, 2020.

[15] Xiaoni Duan, Chien-Ju Ho, and Ming Yin. The influences of task design on crowdsourced judgement: A case study of recidivism risk evaluation. In *The ACM Web Conference (WWW)*, 2022.

[16] Jane Yung-jen Hsu, Kwei-Jay Lin, Tsung-Hsiang Chang, Chien-ju Ho, Han-Shen Huang, and Wan-rong Jih. Parameter learning of personalized trust models in broker-based distributed trust management. *Information Systems Frontiers*, 2006.

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

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

[19] Saumik Narayanan, Guanghui Yu, Chien-Ju Ho, and Ming Yin. How does value similarity affect human reliance in ai-assisted ethical decision making? In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2023.

[20] Saumik Narayanan, Guanghui Yu, Wei Tang, Chien-Ju Ho, and Ming Yin. How does predictive information affect human ethical preferences? In *AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, 2022.

[21] Wei Tang and Chien-Ju Ho. Bandit learning with biased human feedback. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2019.

[22] Wei Tang and Chien-Ju Ho. On the bayesian rational assumption in information design. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2021. **Nominee for Best Paper Award**.

[23] Wei Tang, Chien-Ju Ho, and Yang Liu. Differentially private contextual dynamic pricing. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2020.

[24] Wei Tang, Chien-Ju Ho, and Yang Liu. Optimal query complexity of secure stochastic convex optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[25] Wei Tang, Chien-Ju Ho, and Yang Liu. Bandit learning with delayed impact of actions. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[26] Wei Tang, Chien-Ju Ho, and Yang Liu. Linear models are robust optimal under strategic behavior. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

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

[28] Lauren S Treiman, Chien-Ju Ho, and Wouter Kool. Humans forgo reward to instill fairness into ai. In *AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2023.

[29] Guanghui Yu and Chien-Ju Ho. Environment design for biased decision makers. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2022.

[30] Guanghui Yu, Wei Tang, Saumik Narayanan, and Chien-Ju Ho. Encoding human behavior in information design through deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.