Research Statement: Behavior-Informed Machine Learning

Chien-Ju Ho

December 20, 2023

Machine learning (ML) has integrated into various facets of humans' 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 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 the 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 to 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 key 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. Below I highlight a few of my research projects alone 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, I examined how online workers react to different performance-based payments (Ho et al., 2015). 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. 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 design (Duan et al., 2022) and relaxed the standard data independence assumption by allowing workers to communicate with each other (Duan et al., 2020; Tang et al., 2019).

Another important aspect of human behavior during data collection is humans' awareness of ML. As ML becomes ubiquitous, human behavior might evolve accordingly. For example, if users

are aware their movie ratings are going to impact the movie recommendations they receive in the future, they might update their rating behavior. Together with Psychology researchers, I examined whether human behavior changes when they are aware their behavior will be used to train ML systems (Treiman et al., 2023). Using the classical ultimatum game as the decision-making task, we found that humans are willing to sacrifice their personal gains to improve the fairness of the downstream ML systems when they are aware of the ML training. Moreover, this behavior change is robust whether humans are going to 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, much less attention has been given to intervening in data collection processes. My research has contributed to data-centric ML, focusing on improving the data used to train ML systems. In particular, my earlier works (Ho et al., 2013; Ho and Vaughan, 2012) have explored the problem of assigning labeling tasks to workers and aggregating the obtained labels. Leveraging the online primal-dual techniques, I have developed online 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. The developed algorithms are theoretically shown to achieve near-optimal performance and empirically demonstrated to perform well with real-world crowdsourcing workers. Notably, the online primal-dual techniques developed are general-purpose techniques. I have later applied them to other societal resource allocation problems such as in kidney allocation (Li et al., 2019) and homelessness prevention (Dong et al., 2021).

I have also studied the design of incentives to motivate high-quality data from humans. I explored the problem of learning the optimal performance-based payments, in which workers' payments depend on the quality of their work, in crowdsourcing markets (Ho et al., 2014). I extended the standard principal-agent model from economic theory to a multi-round online model. I designed a novel *bandit algorithm* which only observes limited information from workers but can perform nearly as well as an oracle algorithm which has access to full information. In addition to financial incentives, I have also explore the design of other forms of incentives, such as reputation systems (Ho et al., 2012; Hsu et al., 2006),attention (Liu and Ho, 2018), and social verification (Ho and Chen, 2009). I have also implemented human computation games for collecting data from real-world users in the field (Ho et al., 2007, 2009, 2011).

Accounting for human behavior in machine learning. In addition to understanding human behavior and improve data collection, I have developed learning algorithms to explicitly account for human behavior when learning from human data. My earlier works have focused on the case of strategic human behavior. I explored the problem of actively purchasing data from users for solving machine learning tasks (Abernethy et al., 2015). I showed how to convert a large class of machine learning algorithms into online posted-price and learning mechanisms. The proposed mechanisms identify the *importance* of each data point and decide the payment to offer to each user. I proved that our mechanisms are *incentive-compatible*, i.e., workers are willing to truthfully report their costs. Furthermore, I showed that our mechanisms cost much less while achieving learning accuracies of the same order when compared with purchasing all data points. I also explored the problem of eliciting workers' confidence to achieve optimal label aggregation, with an additional focus on the design of multiple-choice questions (Ho et al., 2016). I developed a Bayesian framework to model the process of eliciting and aggregating data from the crowd. The framework provides an incentive-compatible payment scheme (i.e., workers would truthfully report their confidences), a principled way of aggregating labels, and optimal designs of multiple-choice questions.

Later, I also incorporated psychology-grounded human behavior into machine learning. I have

addressed the problem of bandit learning with biased human feedback (Tang and Ho, 2019), a form of reinforcement learning with human feedback (RLHF). In particular, I consider the setting in which, when eliciting feedback from humans, their feedback is not independently drawn as often assumed in bandit learning. Instead, their feedback is influenced by other users' feedback (also known as herding behavior). By formally incorporating this human behavior into the bandit learning framework, we theoretically demonstrate that under certain mild conditions, we might reach a situation where learning is infeasible, even with an infinite amount of data. This observation reinforces the need for my research in both better understanding human behavior and in improving the 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 (Tang et al., 2021*b*). This approach is applicable to a general set of human behavior models.

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 these decisions. 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, design ML assistance to improve human decision outcomes, and examine the downstream impacts of machine learning.

Understanding human responses to ML assistance. In order to design assistive ML, we need to gain understandings of how humans respond to ML assistance. One framework to address human response to ML assistance is information design (Ding et al., 2023), where humans incorporate ML assistance as new information to update their beliefs about the world, and then make decisions based on the 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. I developed an alternative framework for information design (Tang and Ho, 2021) 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, i.e., when do humans decide to follow ML recommendations. I have conducted a series of human-subject experiments in the context of ethical decision-making (Narayanan et al., 2023, 2022), using kidney allocation as examples. We found that even just the presence of predictive information significantly changes how humans take into account other information and that the source of the predictive information (e.g., whether the predictions are made by ML or humans) plays a key role in how humans incorporate the 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 help improve our understanding of how humans respond to ML assistance, which in turn helps in designing better ML assistance policy.

Designing assistitve ML. My recent works have also started to address the research question of designing ML to assist human decision-making. I investigated the setting in which a (potentially biased) human decision-maker operates in a sequential decision-making environment (Yu and Ho, 2022), and 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) and incorporated common models of biased agents through introducing general time-discounting functions. We then formalized the environment design problem as constrained optimization problems and proposed corresponding algorithms. Our proposed 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 desired outcomes. As highlighted earlier, the vast majority of literature in information design makes strong assumptions about human behavior, and I have proposed an alternative framework with empirically-grounded human behavioral models (Tang and Ho, 2021). Building on top of this framework, I theoretically characterized the (approximately-)optimal information policy within this framework (Feng et al., 2024). Moreover, we also proposed rationality-robust information policies, where the provided information performs well even when we do not have full information on human behavior. While our results have extended information design to settings beyond the standard human rationality assumption, they still only address a subset of alternative human models. I developed a data-driven optimization framework that can work with any provided human models (Yu et al., 2023), including ones where we do not have a closed-form expression of human behavior but have access to human behavioral data. Through extensive simulation, we showed that our data-driven optimization approach not only recovers near-optimal information policies with known analytical solutions, but also can extend to designing information policies for settings that are computationally challenging or for settings where there are no 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, I have examined the long-term impacts of actions in sequential decision-making (Tang et al., 2021*a*). 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 run. It is important to note that this consideration is not solely for promoting fairness; taking into account 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 has formalized the concept of the long-term impact of actions in bandit learning and explored algorithmic designs to help us 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 (Tang et al., 2020*a*,*b*).

My Agenda Going Forward

We have witnessed the rapid growth of the machine learning field, particularly the advent and widespread utilization of generative AI technologies, in the past few years. Correspondingly, the interdependency of human-machine interactions is also rapidly increasing in our daily lives and is set to continue its growth. This escalating interdependency brings into focus the dynamics of human cognition and behavior, especially when individuals interact with machine learning systems or when their behavior influences the training of these systems. Similarly, it's crucial for new ML technologies to account for human behavior and responses to ensure optimal collaboration

between humans and machines. My long-term research agenda is dedicated to advancing our understanding of human behavior in the age of machine learning, and to designing ML-enabled agents that collaborate effectively with humans.

Building on my existing collaborations with psychology researchers (Treiman et al., 2023), one of my research agendas going forward is to empirically examine how the presence of ML changes human behavior and understand the underlying cognitive mechanisms. To conduct the research, following the standard literature, I will start by utilizing social games, such as the ultimatum game, dictator game, and prisoner's dilemma, to examine human behavior with the presence of ML, and then extend the investigations to different domains. These social games provide succinct abstractions of human behavior in different contexts and are useful as the starting point for a comprehensive understanding of humans. I'll examine human behavior and underlying cognitive mechanisms both when explicitly interacting with ML and when implicit interacting with ML through providing training data. The results will not only deepen our understanding of human behavior in the presence of ML but also serve as an improved foundation for learning from behavioral data and assisting human decision-making.

In my existing research, I have investigated various methods for machine learning (ML) to assist humans in decision-making, either by providing assistive information or by modifying the decision-making environment. In these scenarios, ML has predominantly played an assistive role, with humans making the final decisions. However, as ML technology becomes more advanced and widespread, ML systems are increasingly expected to function as teammates, assuming some decision-making responsibilities and working collaboratively with humans in many scenarios. Given this shift, it is crucial to design ML systems capable of anticipating human behavior, enabling them to take actions that complement those of their human partners. Additionally, ML teammates should make their actions and decisions understandable to humans, facilitating effective coordination between human and ML team members. My long-term research agenda is focused on designing collaborative ML agents for human-machine teams. This proposed research revolves around creating virtual AI agents that both comprehend human behavior and make ML behavior comprehensible to humans, ultimately aiming to optimize the performance of human-machine teams. To achieve this, I plan to develop algorithms that enable ML agents to efficiently infer human behavioral models and decision-making characteristics by observing human actions. I also intend to devise methods for ML agents to act in a manner that is intelligible to humans, allowing them to understand the intentions of ML agents. Finally, I will formulate action plans for ML agents that consider both human behavior and beliefs, ensuring optimal performance in human-machine teams.

Lastly, a key objective of mine is to collaborate with domain experts to address practical challenges in deploying my research in real-world domain applications. For example, I plan to adapt my research for application in the field of homelessness prevention, collaborating with Prof. Patrick Fowler at the Brown School of Social Work. In the long term, my goal is to leverage the interdisciplinary resources and expertise available at Washington University (WashU). I plan to expand this research across various application domains by collaborating with several WashU entities. These include the Division of Computational and Data Sciences (DCDS), the Center for Collaborative Human-AI Learning and Operation (HALO), and the Transdisciplinary Institute in Applied Data Sciences (TRIADS). By doing so, I intend to foster a comprehensive, interdisciplinary approach to developing and applying AI in a range of critical and impactful areas.

References

Abernethy, J., Chen, Y., Ho, C.-J. and Waggoner, B. (2015), Cost-efficient learning via active data procurement, *in* 'ACM Conference on Economics and Computation (EC)'.

Ding, B., Feng, Y., Ho, C.-J., Tang, W. and Xu, H. (2023), Competitive information design for pandora's box, *in* 'ACM-SIAM Symposium on Discrete Algorithms (SODA)'.

Dong, Z., Das, S., Fowler, P. and Ho, C.-J. (2021), Efficient nonmyopic online allocation of scarce reusable resources, *in* 'International Conference on Autonomous Agents and Multiagent Systems (AAMAS)'.

Duan, X., Ho, C.-J. and Yin, M. (2020), Does exposure to diverse perspectives mitigate biases in crowdwork? an explorative study, *in* 'AAAI conference on human computation and crowdsourcing (HCOMP)'.

Duan, X., Ho, C.-J. and Yin, M. (2022), The influences of task design on crowdsourced judgement: A case study of recidivism risk evaluation, *in* 'The ACM Web Conference (WWW)'.

Feng, Y., Ho, C.-J. and Tang, W. (2024), Rationality-robust information design: Bayesian persuasion under quantal response, *in* 'ACM-SIAM Symposium on Discrete Algorithms (SODA)'.

Ho, C.-J., Chang, T.-H. and jen Hsu, J. Y. (2007), Photoslap: A multi-player online game for semantic annotation, *in* 'AAAI Conference on Artificial Intelligence (AAAI)'.

Ho, C.-J., Chang, T.-H., Lee, J.-C., jen Hsu, J. Y. and Chen, K.-T. (2009), Kisskissban: A competitive human computation game for image annotation, *in* 'Human Computation Workshop (HCOMP)'.

Ho, C.-J. and Chen, K.-T. (2009), On formal models for social verification, *in* 'Human Computation Workshop (HCOMP)'.

Ho, C.-J., Frongillo, R. and Chen, Y. (2016), Eliciting categorical data for optimal aggregation, *in* 'Advances in Neural Information Processing Systems (NIPS)'.

Ho, C.-J., Jabbari, S. and Vaughan, J. W. (2013), Adaptive task assignment for crowdsourced classification, *in* 'International Conference on Machine Learning (ICML)'.

Ho, C.-J., Slivins, A., Suri, S. and Vaughan, J. W. (2015), Incentivizing high quality crowdwork, *in* 'International World Wide Web Conference (WWW)'. **Nominee for Best Paper Award**.

Ho, C.-J., Slivkins, A. and Vaughan, J. W. (2014), Adaptive contract design for crowdsourcing markets: Bandit algorithms for repeated principal-agent problems, *in* 'ACM Conference on Economics and Computation (EC)'.

Ho, C.-J. and Vaughan, J. W. (2012), Online task assignment in crowdsourcing markets, *in* 'AAAI Conference on Artificial Intelligence (AAAI)'.

Ho, C.-J., Wu, C.-C., Chen, K.-T. and Lei, C.-L. (2011), Deviltyper: A game for captcha usability evaluation, *in* 'ACM Computers in Entertainment'.

Ho, C.-J., Zhang, Y., Vaughan, J. W. and van der Schaar, M. (2012), Towards social norm design for crowdsourcing markets, *in* 'Human Computation Workshop (HCOMP)'.

Hsu, J. Y.-j., Lin, K.-J., Chang, T.-H., Ho, C.-j., Huang, H.-S. and Jih, W.-r. (2006), 'Parameter learning of personalized trust models in broker-based distributed trust management', *Information Systems Frontiers* .

Li, Z., Lieberman, K., Macke, W., Carrillo, S., Ho, C.-J., Wellen, J. and Das, S. (2019), Incorporating compatible pairs in kidney exchange: A dynamic weighted matching model, *in* 'ACM Conference on Economics and Computation (EC)'.

Liu, Y. and Ho, C.-J. (2018), Incentivizing high quality user contributions: New arm generation in bandit learning, *in* 'AAAI Conference on Artificial Intelligence (AAAI)'.

Narayanan, S., Yu, G., Ho, C.-J. and Yin, M. (2023), How does value similarity affect human reliance in ai-assisted ethical decision making?, *in* 'AAAI/ACM Conference on AI, Ethics, and Society (AIES)'.

Narayanan, S., Yu, G., Tang, W., Ho, C.-J. and Yin, M. (2022), How does predictive information affect human ethical preferences?, *in* 'AAAI/ACM Conference on AI, Ethics, and Society (AIES)'.

Tang, W. and Ho, C.-J. (2019), Bandit learning with biased human feedback., *in* 'International Conference on Autonomous Agents and Multiagent Systems (AAMAS)'.

Tang, W. and Ho, C.-J. (2021), On the bayesian rational assumption in information design, *in* 'AAAI Conference on Human Computation and Crowdsourcing (HCOMP)'. **Nominee for Best Paper Award**.

Tang, W., Ho, C.-J. and Liu, Y. (2020*a*), Differentially private contextual dynamic pricing., *in* 'International Conference on Autonomous Agents and Multiagent Systems (AAMAS)'.

Tang, W., Ho, C.-J. and Liu, Y. (2020*b*), Optimal query complexity of secure stochastic convex optimization, *in* 'Advances in Neural Information Processing Systems (NeurIPS)'.

Tang, W., Ho, C.-J. and Liu, Y. (2021*a*), Bandit learning with delayed impact of actions, *in* 'Advances in Neural Information Processing Systems (NeurIPS)'.

Tang, W., Ho, C.-J. and Liu, Y. (2021*b*), Linear models are robust optimal under strategic behavior, *in* 'International Conference on Artificial Intelligence and Statistics (AISTATS)'.

Tang, W., Yin, M. and Ho, C.-J. (2019), Leveraging peer communication to enhance crowdsourcing, *in* 'The World Wide Web Conference (WWW)'.

Treiman, L. S., Ho, C.-J. and Kool, W. (2023), Humans forgo reward to instill fairness into ai, *in* 'AAAI Conference on Human Computation and Crowdsourcing (HCOMP)'.

Yu, G. and Ho, C.-J. (2022), Environment design for biased decision makers, *in* 'International Joint Conference on Artificial Intelligence (IJCAI)'.

Yu, G., Tang, W., Narayanan, S. and Ho, C.-J. (2023), Encoding human behavior in information design through deep learning, *in* 'Advances in Neural Information Processing Systems (NeurIPS)'.