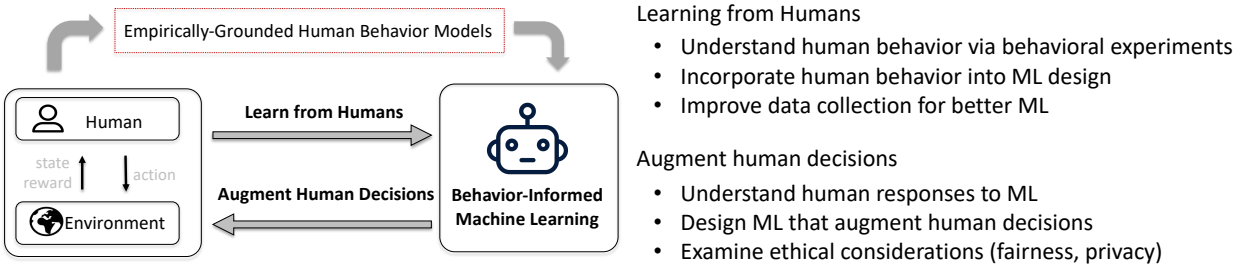
Research Statement: Behavior-Informed Machine Learning

Chien-Ju Ho

December 20, 2023

Machine learning (ML) has integrated into various facets of everyday life, largely deriving its training from human data. Consequently, these ML systems often exhibit and reflect human behavioral biases, leading to concerns in applications from social media to medical decision-making. However, current ML methodologies mostly view humans as independent, stochastic data sources or assume they are rational decision-makers, despite evidence from psychological studies suggesting otherwise. This gap highlights the need to incorporate empirically grounded human behavior insights into ML design. Moreover, as ML continues to grow, it opens up the potential for designing systems that augment human decisions by accounting for realistic human behavior.

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: designing ML systems that learn from human behavioral data, and designing ML systems that augment humans in decision making.



My research investigates interactions between humans and machine learning (ML) and is interdisciplinary in nature. To this end, I have collaborated with and co-advised Ph.D. students alongside faculty from psychology, social work, and biomedical engineering (with a focus on education). My first Ph.D. student, Wei Tang, joined the Business School at the Chinese University of Hong Kong (CUHK) as an assistant professor, demonstrating the interdisciplinary impact of our research. Beyond the general fields of AI and ML, my research contributes to human computation and crowdsourcing by developing theoretically grounded algorithms to leverage crowd data and conducting behavioral experiments to understand the crowd. It also advances the study of multi-armed bandits by incorporating strategic and behavioral human responses and examining their societal impacts. More recently, my work has contributed to understanding and accounting for human behavior in human-AI interactions, particularly in the contexts of information design and environment design. My research has been recognized with best paper nominations at WWW 2015 and HCOMP 2021 and has received support from diverse sources, including NSF, ONR, OpenAI, J.P. Morgan, Amazon, and various seed grants at Washington University.

Learning from Human Behavioral Data

One major line of my research has focused on developing machine learning systems that acquire and learn from human data. This line of work is supported by an NSF/Amazon Fairness in AI grant, an OpenAI Superalignment Fast Grant, a J.P. Morgan Faculty Research Award, a seed grant by WashU OVCR, and a seed grant by WashU TRIADS.

Understanding human behavior through behavioral experiments. Most of the work on the study of ML 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 data generation. For example, I examined how online workers react to financial incentives in the form of performance-based payments [15], a paper nominated for best paper at WWW 2015. By conducting a comprehensive set of experiments with more than 2,000 users, I developed a user behavior model that introduces the concept of *user priors* into the standard economic model. This model is empirically consistent with both our results and prior findings and helps us develop algorithms to elicit high-quality data in another of our works [14]. I have also empirically examined human behavior in other dimensions, including how humans respond to different task designs [5] and their behavior when communicating with others [24, 4].

Another important aspect of human behavior I consider is humans' awareness of ML. As ML becomes increasingly embedded in human decision-making, human behavior may evolve accordingly. Together with Wouter Kool, an assistant professor in Psychology and Brain Sciences, I initiated a research program to examine whether and how humans modify their behavior when they know it will be used to train ML [29, 30]. We found that not only do humans change their behavior while training ML, but these shifts also persist for several days after training ends and may even become habitual. Moreover, this behavioral change occurs regardless of whether individuals will interact with the trained ML in the future. These findings suggest that treating human behavioral data as a black-box dataset can be problematic, as people may adjust their behavior based on their understanding of how their data will be used. Part of my research addresses this challenge by explicitly integrating human behavior into ML design and improving data collection processes and as discussed next.

Incorporating human behavior in ML development. 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. In addition to framing the contract design as a learning problem as discussed above [14], I explored the problem of purchasing data from strategic users for solving ML tasks [1]. I showed how to convert a large class of ML algorithms into online posted-price and learning mechanisms. These mechanisms identify the *importance* of each data point to help determine the data pricing. Our mechanisms are *incentive-compatible* and cost significantly less while achieving accuracies of the same order as purchasing all data points. Similarly, I also explored the problem of eliciting workers' confidence in an incentive-compatible manner to achieve optimal label aggregation, with an additional focus on the design of multiple-choice questions [16].

In my more recent research, I incorporated psychology-grounded human behavior into ML. I investigated bandit learning with biased human feedback [22], a form of reinforcement learning with human feedback (RLHF). In particular, instead of assuming human feedback is independent of others, as is commonly done in the literature, I address cases where human feedback might be influenced by other users' feedback (also known as herding behavior). In addition to designing learning algorithms to account for this human bias when learning is feasible, notably, I demonstrate

that under certain mild conditions, learning may not even be feasible—even with an infinite amount of data. This finding underscores the need for my research in both better understanding human behavior and improving data collection. Beyond incorporating specific human behaviors, I have also leveraged robust optimization to design decision rules that remain effective in situations where human models are unknown a priori [28]. Moreover, I have addressed situations that humans might change their behavior in response to ML deployment and account for this potential behavioral shift in ML development.

Improving data collection: Towards data-centric ML. As highlighted in my research, human behavior influences the generation of data used to train ML and may even render downstream ML training infeasible. Therefore, in addition to understanding and incorporating human behavior into ML development, I also devote significant effort to investigating ways to collect better data from humans in the first place. More broadly, my research has contributed to data-centric ML, particularly in improving crowdsourced data collection. My earlier works [8, 13] have explored the problem of task assignment and label inference in crowdsourced data collection. Leveraging online primal-dual techniques, I have developed algorithms that are theoretically proven to be near-optimal and empirically validated to perform well in practice. Notably, the online primal-dual techniques developed are general-purpose techniques. I have later also applied them to other societal resource allocation problems such as kidney allocation [18] and homelessness prevention [3].

I have also studied the problem of incentive design to solicit high-quality data from humans. I explored the problem of learning the optimal crowdsourcing contracts, in which workers' payments depend on the quality of their work [14]. I extended the standard principal-agent model to a multi-round online model and designed a novel *bandit algorithm* which, despite only observing limited information from workers, can perform nearly as well as an oracle algorithm with access to full information. In addition to designing financial incentives, I have also explored the design of other forms of incentives, such as reputation systems [12, 17], attention [19], and social verification [7]. Furthermore, I have explored the idea of using fun as an incentive through implemented human computation games for collecting data from real-world users in the field [9, 10, 11].

More recently, I have examined how to leverage task design to improve crowdsourced data. By utilizing de-biasing strategies from psychology and enabling worker communication, I designed intervention mechanisms during the data annotation process to improve the outcomes of crowdsourced data annotation [24, 4, 5]. Moreover, given that many recent concerns about ML stem from the (lack of) representativeness in training datasets, my ongoing work has explored adaptive data acquisition to improve data representativeness and examined its impact on the fairness of downstream ML models. I have also investigated whether approaches such as perspective-taking can be leveraged to collect more representative data for subjective annotation tasks.

Augmenting Humans in Decision Making

As ML grows more powerful, it opens up the rich potential to design ML systems to augment and assist human decisions. Recently, I have started to investigate approaches to achieve this objective, including understanding human decision-making with ML assistance, designing ML assistance to improve human decision outcomes, and examining the downstream ethical impacts of machine learning. This line of work is supported by a grant from the Office of Naval Research (ONR), an NSF/Amazon Fairness in AI grant, a second J.P. Morgan Faculty Research Award, a second seed grant from WashU TRIADS, and a seed grant from the McDonnell International Scholars Academy.

Understanding human responses to ML assistance. In order to design ML to augment humans, we first need a better understanding of how humans respond to ML assistance. To this

end, I have investigated how humans incorporate information in decision-making. I extended the framework of information design [2] in economics, where humans are often assumed to be Bayesian rational. To relax the strong assumption on human rationality, I developed an alternative framework [23] based on the discrete choice model and probability weighting, a paper nominated as a best paper in HCOMP 2021. Through behavioral experiments, I demonstrated that our framework better explains real-world user behavior. Moreover, building on this framework, in my later works, I have investigated the theoretical characterization of the optimal policy to design information [6] and designed a data-driven optimization framework for finding the optimal policy [32].

I have also examined factors that impact humans' trust and reliance on ML assistance. I have conducted a series of human-subject experiments in the context of ethical decision-making [20, 21]. We found that the mere 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 key role in how humans incorporate the information. Moreover, when humans and ML recommendations disagree, humans are more likely to change their opinion if the ML displays similar *values* to human decision-makers. These projects help improve our understanding of how humans respond to ML assistance, which in turn aids in designing better ML assistance.

Designing ML to augment human decisions. In addition to understanding human responses to ML, my recent works have also started to address the research question of designing ML to augment human decision-making, taking into account human behavior. For example, I have investigated the setting in which a (potentially biased) human decision-maker operates in a sequential decision-making environment [31], and our goal is to design ways to update the decision-making environment or providing action recommendations to improve the human decision outcome. I formalized this *environment design* problem as a constrained optimization problem. I showed that this optimization problem is generally NP-hard and provided efficient algorithms for a relaxed formulation. To showcase the effectiveness of the proposed approach, I have conducted human-subject experiments and demonstrated that our approach can indeed lead to environments that improve human decision outcomes.

Another approach to improve human decisions with ML assistance is through designing assistive information provided to humans. To this end, I have investigated the problem of information design with realistic human behavior. Building on our own empirically-grounded framework [27], I have theoretically characterized the (approximately) optimal information policy within this framework [6]. Moreover, I also proposed rationality-robust information policies to address the common situation where the exact human behavior is not available. In addition to theoretical results, in another work, I developed a data-driven optimization framework that can work with any provided human models [32], including ones where we do not have a closed-form expression but only have access to human behavioral data. Again, through simulations and human-subject experiments, the proposed approach is shown to capture human behavior from data and lead to more effective information policies for real-world human decision-makers. While the current research progress has been more conceptual, I am now working on research projects in adapting our findings in the context of *explanation design* in AI-assisted decision making.

Ethical considerations. When leveraging ML for decision making, it is important to understand its consequences. I have investigated various ethical considerations related to deploying machine learning algorithms in societal domains. As one prominent example, I examined the long-term impacts of actions in sequential decision-making [27]. Take loan applications as an example: 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. I have 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 [25, 26] and understanding human perceptions of fairness in ML deployment.

My Agenda Going Forward

We have witnessed the rapid growth of ML, particularly the advent and widespread utilization of generative AI technologies, in the past few years. Correspondingly, the interactions between humans and ML are also rapidly increasing and are set to continue their growth. It is more crucial than ever to develop ML systems that account for human behavior and responses to ensure effective collaboration between humans and machines. My long-term research agenda is dedicated to advancing our understanding of human behavior in the age of ML and designing ML systems that collaborate effectively with humans.

One of my research goals going forward is to examine how the growing presence of ML in our decision-making pipeline changes human behavior and to understand the underlying cognitive mechanisms. This newly initiated research program is in collaboration with Wouter Kool, an assistant professor in psychology, with whom I am currently co-advising a Ph.D. student. I will first utilize social games to examine human behavior both in the presence of ML and in anticipation of ML, and then extend the investigations to real-world domains. I will study human behavior and underlying cognitive mechanisms both when explicitly interacting with ML and when implicitly interacting with ML through providing training data. The results will not only deepen our understanding of human behavior when ML is incorporated into the decision loop but also serve as an improved foundation for learning from behavioral data and assisting human decision-making.

Another research objective I plan to push forward is to design ML agents that can work alongside humans as teammates. In my current research, I have investigated various methods for ML to assist humans, while humans are still the final decision-makers. However, as ML becomes more advanced and widespread, ML systems are increasingly expected to function as teammates, assuming some decision-making responsibilities and working collaboratively with humans. There are a couple of research directions I hope to pursue along this line. First, with the recent support received from OpenAI, I plan to develop ML agents that behave like humans while achieving higher task performance. Developing human-like ML agents has the potential to lead to more effective interactions with humans, both in the context of collaboration and education. I will develop approaches to investigate this direction. Second, I plan to investigate the design of ML teammates, such as robots, that work alongside humans. To achieve this goal, we need to design ML systems that are capable of both anticipating human behavior and making their actions and decisions understandable to humans. I intend to collaborate with William Yeoh, an expert in planning, and Ioannis Kantaros, an expert in robotics, to develop these ML teammates.

Lastly, I will collaborate with domain experts to address practical challenges in deploying my research in real-world applications. In particular, I will work with Patrick Fowler, an associate professor in Social Work, to explore the deployment of my research in the domain of housing security. I will also work with Dennis Barbour, a professor in Biomedical Engineering, to connect our research findings in the domain of education. I am currently co-advising Ph.D. students with both of them. In the long term, my goal is to leverage the interdisciplinary resources and expertise available at Washington University, including 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

[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] 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.

[3] 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.

[4] Xiaoni Duan, Zhuoyan Li, 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.

[5] 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.

[6] Yiding Feng, Chien-Ju Ho, and Wei Tang. Rationality-robust information design: Bayesian persuasion under quantal response. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2024.

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

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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**.

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

[17] 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.

[18] 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.

[19] 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.

[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] 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.

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

[23] 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**.

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

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

[26] 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.

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

[28] 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.

[29] 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.

[30] Lauren S Treiman, Chien-Ju Ho, and Wouter Kool. The consequences of AI training on human decision making. *Proceedings of the National Academy of Sciences (PNAS)*, 2024.

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

[32] 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.