Design Principles for Building Robust Human-Robot Interaction Machine Learning Models

Josh Bhagat Smith
Oregon State University
Corvallis, OR
bhagatsj@oregonstate.edu

Vivek Mallampati
Oregon State University
Corvallis, OR
mallampv@oregonstate.edu

Prakash Baskaran
Oregon State University
Corvallis, OR
baskarap@oregonstate.edu

Mark-Robin Giolando
Oregon State University
Corvallis, OR
giolandm@oregonstate.edu

Julie A. Adams
Oregon State University
Corvallis, OR
adamsjuli@oregonstate.edu

ABSTRACT

Effective collaboration between humans and robots hinges on the robot's ability to comprehend its human teammate. This collaboration demands the development of machine learning models that bridge the gap between human physiological signals and their mental states. However, the challenge lies in developing generalizable machine learning models using data collected in controlled experimental conditions. This manuscript proposes a set of principles for designing human subject evaluations, emphasizing the crucial balance between experimental control and ecological validity while also balancing fundamental machine learning trade-offs.

CCS CONCEPTS

• **Human-centered computing → HCI design and evaluation methods**.

KEYWORDS

design principles, machine learning, human-robot interaction

ACM Reference Format:
Josh Bhagat Smith, Vivek Mallampati, Prakash Baskaran, Mark-Robin Giolando, and Julie A. Adams. 2024. Design Principles for Building Robust Human-Robot Interaction Machine Learning Models. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24 Companion), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3610978.3640598

1 INTRODUCTION

Successful human-robot teaming consists of humans and robots collaborating to achieve tasks in uncertain, dynamic environments. Deploying robots alongside humans in these environments will require the robots to have a robust and dynamic understanding of their human teammates. Endowing the robot with these capabilities is most commonly accomplished by developing machine learning models. For example, workload estimation models and task recognition models learn the relationships between the human's physiological signals and their mental and physical state.

Collecting data to train these models requires conducting human subject evaluations that mimic the desired applications domain (e.g., disaster response). Many evaluations are conducted under tightly controlled experimental conditions to ensure the precise manipulation of independent variables. However, prioritizing experimental constraints over ecological validity inevitably results in machine learning models that fail to generalize in real world settings. This lack of generalizability is the result of training data that is not fully representative of real world conditions.

Designing human subject evaluations such that resulting machine learning models generalize effectively requires carefully analyzing the machine learning and human-robot interaction (HRI) considerations that are often ignored for the sake of experimental control. These considerations are discussed and principles for designing HRI domain machine learning models that generalize more effectively are presented. These principles highlight the importance of prioritizing the ecological validity of human-robot teaming dynamics and experimental conditions, while also balancing key machine learning trade-offs. These principles also enable experimenters to better understand the extent to which their models generalize (i.e., performance in the real world) and the way in which their models generalize (e.g., across individuals, across tasks).

2 RELATED WORK

A major takeaway from the machine learning community is the impact of data quality and modality on machine learning models [20]. Acknowledging these factors is imperative to draw meaningful conclusions and to understand how models can be deployed. These factors are particularly important when building machine learning models of human behavior, as these models can only be developed using data from human subject evaluations (i.e., small datasets). Experimental design has been extensively explored in structured environments, where the primary focus has been on evaluating the HRI dynamics under specific experimental conditions [11, 13]. Prior work emphasized the limitations of relying on a single evaluation method, advocating for integrating diverse methods (e.g., psychological measures, performance metrics, and behavioral measures) [3]; however, little work has been done to understand how these factors impact resulting machine learning models.

HRI '24 Companion, March 11–14, 2024, Boulder, CO, USA

ACM ISBN 979-8-4007-0323-2/24/03.
https://doi.org/10.1145/3610978.3640598

Successfully developed workload estimation [4] and emotion recognition models [22] have drawn attention to the importance of recognizing physiological changes in the human's mental states. These approaches are sensitive to sensor noise and the variability of human physiology [5]. Constructing models is also heavily influenced by the inconsistencies between experimental conditions and the real world. Task design is critical to ensure tasks are ecologically valid. The task complexity adds another layer of consideration, emphasizing the necessity of clearly defining the context to enhance the realism and relevance of interaction scenarios [14]. This synthesis of insights from the literature sets the stage for a more holistic approach to designing HRI evaluations, aiming to capture complex human-robot interactions in varied and realistic settings.

3 MOTIVATION

Enhancing a robot's capacity to estimate different aspects of its human teammates (e.g., task execution, situational awareness, workload, affect), such that it can make predictions or adaptations to accommodate that teammate, will enhance the efficacy of human-robot teams in real world problem domains. Developing machine learning models capable of modeling these aspects is inherently complex due to the natural variability of human behavior [27]. Further, collecting data sufficiently representative of these behaviors is also difficult. The experimental design must consider how emphasizing ecological validity impacts the experimental design (e.g., data collection, human-robot teaming dynamics), key machine learning algorithm choices, and resulting model robustness.

3.1 Experimental Design Considerations

Achieving real world generalizability with HRI machine learning models requires evaluations that adequately reflect the dynamic, uncertain nature of the real world while maintaining sufficient experimental control to properly manipulate independent variables. Striking this balance is a challenge that requires considering ecologically valid human-robot teaming dynamics, sufficient task diversity, and appropriate data collection procedures.

Ecological validity refers to the degree to which an experiment can be used to predict behaviors in real world settings [13]. There are many different considerations experimenters must make to ensure the ecological validity of their evaluations. Ecological validity for human-robot teams is achieved by ensuring sufficient realism for all human-robot teaming dynamics. Enumerating the full spectrum of human-robot teaming dynamics that must be considered is outside the scope of this manuscript, but a few are important to highlight. First, the division of labor and interactions between the human and the robot must reflect real world human-robot teams. Properly designed robot roles are especially difficult to achieve in Wizard-of-Oz studies [21], as the remote operator must strive to be consistent with how the robot is piloted and avoid misrepresenting the robot's real world function. Second, tasks performed by the human-robot team must be reflective of real world conditions. Task realism is critical to ensuring that the experiment will evoke the corresponding behavior and performance, such that the data collected is useful for developing the machine learning model.

An equally important component is the task diversity. Conducting an evaluation with a single ecologically valid task constrains the conditions in which it is appropriate to deploy the resulting machine learning model. Human-robot teams perform a wide range of tasks that vary based on task coupling, duration, team expertise. HRI evaluations must capture, at least in part, the broad spectrum of real world tasks to successfully develop machine learning models.

Complex HRI evaluations must also consider how these machine-learning models will be deployed. Specifically, collecting data in experimental conditions must not rely on sensors or systems that cannot be practically deployed in realistic environments. Workload estimation and task recognition algorithms map physiological signals to the corresponding aspects of the human's state. Measuring these signals with static sensors (e.g., cameras), or sensors susceptible environmental noises (e.g., EEG [25]), prevents deploying these algorithms in unstructured, real world environments; thus, wearable sensors must be utilized. Real world considerations like these must be made when designing experiments.

3.2 Machine Learning Considerations

Collecting enough data to capture the full range of human behavior is intractable; thus, HRI machine learning models must be constructed using relatively small datasets. These constraints constitute a low data regime and make the application of larger machine learning models impractical [7, 15]. Small datasets are rarely fully representative of the true distribution for a given problem domain, especially when collected in tightly controlled experimental environments. There are three key machine learning issues that must be considered when learning with small datasets: Long-tail distributions, Out-of-distribution (OOD) data and evaluation overfit.

A common problem with task recognition models is long-tailed data distributions. These distributions are characterized by certain tasks that account for the vast majority of data (e.g., $\geq 80\%$ [28]); thus, key tasks are under represented within the dataset. Training machine learning algorithms on such imbalanced datasets will bias the algorithm and lead to performance degradation [16]. Conducting a task analysis [27] of human-robot teaming tasks often reveals that there are many subtasks that occur infrequently, especially in highly dynamic and uncertain real world domains, such as disaster response. Designing experiments such that the ecological validity of these tasks is maintained, but sufficient data is generated to mitigate this natural imbalance is a challenge.

Data that is characterized by a different distribution than the training datasets is considered to be OOD [24]. Enumerating the broad scope of ways that a distribution can change such that the machine learning model is affected is outside the scope of this manuscript [19]; however, this problem is summarized. Broadly, machine learning is the process of learning a function, $f(x)$, that maps from a domain (i.e., features), X, to a range (i.e., labels), Y. Any meaningful changes to either X, Y, or $f(x)$ between the training dataset and the testing dataset constitute a meaningful distribution shift. These problems are prevalent in HRI application domains. One example is training a model on convenience participants and deploying the model on expert users. Another example is training a model on data collected from a single human, single robot team and deploying it on a single-human, multi-robot team. Understanding the impact of these differences and developing models that are robust to these changes is an active area of research.

Evaluation overfit simply means accidentally overfitting to the unseen biases within the dataset. An evaluation's tasks may be ecologically valid, but intrinsic aspects of those tasks (e.g., robot's voice, capabilities) may bias participant behavior in unexpected ways. These biases may not be noticeable in cross-validation procedures, such as leave-one-subject-out [10], as it is present in all data points. Additional validation is necessary to ensure these issues are avoided. All three of these issues are exasperated by humans' individual differences. Fundamentally, human-centric evaluations are noisy. Individual humans have variable physiological responses to a task's demands and may employ different strategies to complete that task. These individual differences make building machine learning models of human behaviors and states for real world problem domains difficult. Additional considerations need to be made for increased sensor noise over time due to human factors (e.g., fatigue), sensor issues (e.g., slippage, drift), and environmental changes.

4 PRINCIPLES

Designing complex HRI evaluations that can be used to train machine learning models for real world application domains requires a broad range of experimental and machine learning considerations. The experiment itself must be adequately representative of the application domain, practical considerations (e.g., sensor deployment) must be considered upfront, and the extent to which the machine learning models can applied must be thoroughly understood.

4.1 Principle 1: Ecological Validity

Ecological validity refers to the real world generalizability of an evaluation (e.g., tasks, interactions, robot's capabilities, form [13]). Overly constrained experiments create artificial human-robot teaming dynamics, which hinders real world generalization and introduces bias into the machine learning models. The following subset of experimental design factors must be considered in order to achieve ecologically valid human-robot teaming dynamics.

Real world human-robot teams can be deployed for long periods of time (i.e., hours, days). Data based on short-duration tasks (e.g., ten minutes) does not reflect a human's behavior and perception changes over the longer-duration real world tasks. An evaluation's ecological validity and machine learning models' generalizability are heavily impacted by evaluation tasks' durations.

Task density is a common variable to modulate workload [26]; however, task densities necessary to manipulate the evaluation's independent variables may rarely be encountered in the real world. These uncommon task densities may impact human behavior, and interactions with robots may bias the underlying patterns the machine learning model discovers and hinder generalizability.

Ensuring that a fully autonomous robot can interact with a human in a reliably safe manner is challenging. The Wizard-of-Oz experimental technique [21] allows experimenters to remotely operate the robot as needed to ensure safe and consistent behavior, as well as respond to any unforeseen scenarios (e.g., maneuvering the robot during navigational failures). However, operating the robot in a human-like manner may compromise the evaluation's ecological validity. Developing specific criteria for how the human operates the robot, such that it mimics autonomous behaviors, are essential when using Wizard-of-Oz techniques; otherwise, machine learning models may learn human-robot teaming dynamics uncharacteristic of real world teams with autonomous robots.

The robot's capabilities and physical form significantly alter the tasks and interactions that can be performed. The tasks and interactions directly inform the human's response to the evaluation's demands, which can artificially alter the human's behavior. These behavioral changes may constitute a meaningful difference in the resulting machine-learning model. Furthermore, the team composition (i.e., the number of robots and humans) significantly impacts human-robot teaming dynamics. Gathering data for different team compositions may constitute a meaningful difference, preventing the generalization of the machine learning model.

Evaluating the ecological validity of the experimental design factors is critical to developing generalizable machine learning models; however, this list is not comprehensive. Ensuring that all environmental conditions, human-robot teaming dynamics, and participant-experiment interactions are grounded in the intended application domain is central to minimizing the artificial aspects of an evaluation; thus, maximizing ecological validity.

4.2 Principle 2: Variety in Tasks and Team Dynamics

Machine learning models are only as expressive as the datasets used to train them. Ecologically valid evaluations ensure that data generated is representative of real world human-robot teaming, but machine learning models trained on a subset of tasks are likely to overfit to specific aspects of those tasks. Therefore, evaluations must be designed to reflect the real world variability of a given application domain. Designing experiments that capture this diversity must abide by strict principles to ensure that Principle 1 is not violated. First, an evaluation must have clear defined HRI roles (e.g., supervisor, peer) assigned to all humans operating in the team. Second, the evaluation must ensure that the tasks executed by a team are appropriately grounded in the respective HRI roles and that tasks represent a given application domain's diversity.

Generally, human-robot teaming diversity is equally important as task diversity. For example, a peer-based human-robot team must accomplish a variety of tasks that are independent, loosely-coupled, and tightly-coupled [18]. Machine learning models that make accurate estimates across these scenarios must be trained on data that fully captures these dynamics. Capturing this variability is best accomplished through domain expertise or collaboration with domain experts, as selecting tasks that fully represent the human-robot teaming dynamics is difficult and varies across domains.

An important teaming aspect for complex HRI evaluations is an evaluation's task duration. Most evaluations consists of artificially short tasks and it is difficult to capture meaningful HRI data when interactions last less than a minute. Experiments must incorporate longer tasks, or longer time periods consisting of a sequence of short duration tasks, to be representative of the real world. Models built to estimate such human-based responses need to account for these complexities to afford application to real world domains.

4.3 Principle 3: Sensor Suite Selection

Collecting data across the diverse tasks performed by human-robot teams hinges on selecting the appropriate set of sensors that can

accurately capture the human's interactions, but can also be successfully deployed in unstructured, dynamic environments. Sensors that restrict the human-robot team's interactions spatially (e.g., environmentally embedded cameras), temporally (e.g., sensors with low battery life), or computationally (e.g., require high storage) must be avoided when possible and viable alternatives must be considered. For example, using whole-body IMU-based motion tracking, instead of optical motion capture systems, allows participants to move freely. Sensors used in experimental conditions to collect training data must be easily deployed in realistic settings; otherwise, the resulting machine learning cannot be used.

Many wearable sensors exhibit increased sensor noise and variability over their non-wearable counterparts (e.g., stationary vs. mobile eye-trackers [8]). Using multiple sensors is necessary to prevent machine learning models from over relying on a single noisy sensor. Further, sensors must capture the full spectrum of human activity so that the developed machine learning models can be used in diverse scenarios. Sensors required to asses the human's state for cognitive tasks are substantially different than the sensors required for physical tasks [1]. Sensor suites allow machine learning models to learn non-trivial interactions between multiple aspects of the human's mental and physical state. Experimenters must considered the real world constraints of utilizing a particular sensor suite and the extent to which that sensor suite fully captures the desired aspect of the human's behavior.

4.4 Principle 4: Robust Model Validation

Deploying machine learning models of human behavior in realistic settings requires a comprehensive understanding of the circumstances in which these models perform well, which is achieved through robust model validation. Prior work designed human subject evaluations to build machine learning models capable of estimating latent properties of the human's internal state (e.g., workload [5], situational awareness [12]). These latent properties are not directly observable; therefore, they are difficult to verify. Ensuring that the machine learning models' output accurately reflects a human's internal state is paramount to the successful deployment of machine learning models in real world human-robot teams.

Ground truth values collected during an evaluation of the models must be derived from objective sources whenever possible. Relying on subjective questionnaires to develop models automatically encodes the reporting errors and inherent biases [17]. These errors and biases introduce additional noise into the machine learning model's training process, making it more difficult to learn the relationship between the external aspect (e.g., physiological signals) and latent properties of the human's state.

Additionally, validating a machine learning model using standard techniques [23] (e.g., leave-one-subject-out) only measures the model's ability to generalize within the evaluation's context. These techniques do not speak to the model's utility in unknown scenarios. Specifically, evaluations may possess unexpected biases inherent in the experimental conditions (e.g., tasks, human-robot interactions), which machine learning models will inadvertently learn. Validating machine learning models in a separate evaluation, or in a real world deployment, helps experimenters verify the extent which the machine learning model can generalize and ensures that the model did not unintentionally learn these inherent biases.

4.5 Principle 5: Dataset Composition Trade-Offs

Effective use of machine learning models in real world HRI domains requires flexible models that account for human's individual differences. The dynamics of task execution vary day-to-day, between individuals, and are influenced by external factors. Machine learning models developed in rigid experimental environments are unlikely to account for these individual differences due to inter-day and environmental variances or differences across individuals. Real world problem domains are inherently imbalanced, as some tasks naturally occur more frequently than others. Datasets that exhibit this level of variability and imbalance will exhibit either i) a long-tail distribution, where certain classes (e.g., tasks) account for the vast majority of the data, while all the other classes are underrepresented, or ii) OOD data, where the training and testing data are drawn from different distributions, even though the context, task objective, and environment remain fairly similar.

Overcoming these challenges can be solved experimentally or algorithmically, but with trade-offs. For example, an experimental solution to long-tailed distributions explicitly incorporates infrequent tasks, which may result in compromising ecological validity. Downsampling overrepresented classes [6] and weighting underrepresented classes [9] are both algorithmic solutions that preserve ecological validity, but neither solution is guaranteed to overcome performance degradation due to class imbalances. Likewise, experimentally incorporating untrained tasks (i.e., OOD data) for the purposes of using non-standard machine learning methods comes with similar trade-offs. The degree of difference between trained and untrained tasks directly impacts the model's ability to scale to untrained or novel real world tasks [2]. These trade-offs must be considered, and the corresponding decision must be explicitly discussed in order for future applications to understand the circumstances in which the machine learning models will generalize.

5 CONCLUSION

Principles for designing evaluations that capture the complex dynamics of human-robot teams and result in generalizable machine learning models were presented. Complex HRI evaluations are required for building machine learning models that enable a robot to understand a human's state, such that the two can effectively collaborate in dynamic, uncertain domains. These principles are an initial attempt at enumerating means of addressing these difficulties.

6 ACKNOWLEDGEMENTS

The students were supported University Space Research Association contract 904186092, and ONR grants N00024-20-F-8705 and N00014-21-1-2052. The contents are those of the authors and do not represent the official views of, nor an endorsement, by research sponsors, or the U.S. Government.

REFERENCES

[1] P. Baskaran and J. A. Adams. Multi-dimensional task recognition for human-robot teaming: literature review. *Frontiers in Robotics and AI*, 10, 2023.

[2] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. W. Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

[3] C. L. Bethel and R. R. Murphy. Review of human studies methods in HRI and recommendations. *International Journal of Social Robotics*, 2(4):347–359, 2010.

[4] J. Bhagat Smith, P. Baskaran, and J. A. Adams. Decomposed physical workload estimation for human-robot teams. In *International Conference on Human-Machine Systems*, pages 1–6. IEEE, 2022.

[5] J. Bhagat Smith, S. A. Toribio, P. Baskaran, and J. A. Adams. Uncertainty-aware visual workload estimation for human-robot teams. In *Conference on Cognitive and Computational Aspects of Situation Management*, pages 1–8, 2023.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[7] X.-W. Chen and X. Lin. Big data deep learning: challenges and perspectives. *IEEE Access*, 2:514–525, 2014.

[8] S. Dowiasch, P. Wolf, and F. Bremmer. Quantitative comparison of a mobile and a stationary video-based eye-tracker. *Behavior research methods*, 52:667–680, 2020.

[9] C. Elkan. The foundations of cost-sensitive learning. In *International Joint Conference on Artificial Intelligence*, volume 17, pages 973–978. Lawrence Erlbaum Associates Ltd, 2001.

[10] M. Esterman, B. J. Tamber-Rosenau, Y.-C. Chiu, and S. Yantis. Avoiding non-independence in fMRI data analysis: Leave one subject out. *Neuroimage*, 50(2):572–576, 2010.

[11] M. R. Fraune, I. Leite, N. Karatas, A. Amirova, A. Legeleux, A. Sandygulova, A. Neerincx, G. Dilip Tikas, H. Gunes, M. Mohan, N. I. Abbasi, S. Shenoy, B. Scassellati, E. J. de Visser, and T. Komatsu. Lessons learned about designing and conducting studies from HRI experts. *Frontiers in Robotics and AI*, 8:772141, 2021.

[12] M. Ginesi, D. Meli, A. Roberti, N. Sansonetto, and P. Fiorini. Autonomous task planning and situation awareness in robotic surgery. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3144–3150. IEEE, 2020.

[13] G. Hoffman and X. Zhao. A primer for conducting experiments in human–robot interaction. *ACM Transactions on Human-Robot Interaction*, 10(1), 2020.

[14] G. A. Holleman, I. T. C. Hooge, C. Kemner, and R. S. Hessels. The 'real-world approach' and its problems: A critique of the term ecological validity. *Frontiers in Psychology*, 11:721, 2020.

[15] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[16] A. Luque, A. Carrasco, A. Martín, and A. de Las Heras. The impact of class imbalance in classification performance metrics based on the binary confusion matrix. *Pattern Recognition*, 91:216–231, 2019.

[17] G. Matthews, J. De Winter, and P. A. Hancock. What do subjective workload scales really measure? operational and representational solutions to divergence of workload measures. *Theoretical Issues in Ergonomics Science*, 21(4):369–396, 2020.

[18] L. Mingyue Ma, T. Fong, M. J. Micire, Y. K. Kim, and K. Feigh. Human-robot teaming: Concepts and components for design. In *Conference on Field and Service Robotics*, pages 649–663. Springer, 2018.

[19] K. Murphy. *Probabilistic Machine Learning: Advanced Topics*. MIT Press, Cambridge, MA, 2023.

[20] S. Rangineni. An analysis of data quality requirements for machine learning development pipelines frameworks. *International Journal of Computer Trends and Technology*, 71, 2023.

[21] L. D. Riek. Wizard of oz studies in hri: A systematic review and new reporting guidelines. *Journal of Human-Robot Interaction*, 1(1):119–136, 2012.

[22] P. Schmidt, A. Reiss, R. Dürichen, and K. V. Laerhoven. Wearable-based affect recognition—a review. *Sensors*, 19(19):4079, 2019.

[23] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson. Machine learning algorithm validation with a limited sample size. *PloS one*, 14(11):e0224365, 2019.

[24] J. Wang, C. Lan, C. Liu, Y. Ouyang, T. Qin, W. Lu, Y. Chen, W. Zeng, and P. Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[25] E. Wascher, J. Reiser, G. Rinkenauer, M. Larrá, F. A. Dreger, D. Schneider, M. Karthaus, S. Getzmann, M. Gutberlet, and S. Arnau. Neuroergonomics on the go: An evaluation of the potential of mobile EEG for workplace assessment and design. *Human Factors*, 65(1):86–106, 2023.

[26] M. B. Weinger, O. W. Herndon, M. H. Zornow, M. P. Paulus, D. M. Gaba, and L. T. Dallen. An objective methodology for task analysis and workload assessment in anesthesia providers. *Anesthesiology*, 80(1):77–92, 1994.

[27] C. D. Wickens, S. E. Gordon, Y. Liu, and J. Lee. *An Introduction to Human Factors Engineering*, volume 2. Pearson Prentice Hall Upper Saddle River, NJ, 2004.

[28] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng. Deep long-tailed learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10795–10816, 2023.