Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI-24)
Early Career Track

Human-Robot Alignment through Interactivity and Interpretability: Don't Assume a "Spherical Human"

Matthew Gombolay
Georgia Institute of Technology
Matthew.Gombolay@cc.gatech.edu

Abstract

Interactive and interpretable robot learning can help to democratize robots, placing the power of assistive robotic systems in the hands of end-users. While machine learning-based approaches to robotics have achieved impressive results, robot learning is still a feat of costly engineering performed in controlled settings and relying upon impractical assumptions about humans. To achieve a vision in which robots can be integrated sustainably into our daily lives for robotic assistance, researchers must take a human-centered approach and develop novel approaches for human-robot alignment of robot values and behaviors. This paper amalgamates recent human factors insights and computational techniques that can support human-robot alignment through interactive and interpretable robot learning and teaming.

1 Introduction

Robots have the potential to support every aspect of our daily lives, beyond the dull, dirty, and dangerous (3D) tasks roboticists have traditionally focused on [Takayama *et al.*, 2008]. One such area is in supporting adults with mild cognitive impairment (MCI) who often need nursing care and to live with a care partner who manages their activities of daily living (ADL). Yet the cost of at-home nursing has risen to exceed household disposable income [PRNewswire, 2016; Prince and Fantom, 2014], and there is a worsening nursing shortage [Institute of Medicine, 2010]. Labor shortages will impact almost every sector due to decreased human fertility globally [Skakkebæk *et al.*, 2022]; while robots may not be a perfect solution today [Wright, 2023], they may be a necessary boon. Beyond the crises on Earth, robots may be the primary avenue we have to explore the solar system [Jayanthi *et al.*, 2023] or galaxy due to our fragile, finite bodies, the vastness of space, and the cosmic speed limit, c.

Traditionally, developing and deploying robots has required an army of consultants to program and install fixed robotic systems that are expensive and not easily adapted to dynamic user needs. This approach has only been successful in industries where economies of scale and relatively static, standardized product lines were available, such as in automotive manufacturing. Further, these robots are typically caged off from human workers due to their lack of safety, adaptability, and interactivity. However, the ubiquitous robots envisioned in this paper need a scalable approach that goes beyond these hand-crafted, expert systems.

Researchers have made tremendous progress in automating the process of developing robot controllers through Reinforcement Learning (RL) but have not solved the problem of removing the human expert. Unfortunately, applying RL to real-world scenarios typically relies upon extensive, human reward engineering, hyperparameter tuning, and trial-and-error design of ad hoc neural network architectures that cannot generalize across tasks and domains. Fundamentally, these classes of approaches attempt to learn *tabula rasa*, and we lack informative priors on robot exploration by which robots can quickly synthesize new control policies without human intervention.

In contrast, the field of *interactive robot learning*[Seraj *et al.*, 2024] seeks to get the best of both worlds between RL and expert systems by enabling robots to learn from human-robot interaction (HRI), including human task demonstrations (i.e., Learning from Demonstration (LfD)) [Chen *et al.*, 2021], natural language instruction and feedback [Silva *et al.*, 2021; Tambwekar *et al.*, 2023], accessible programming interfaces [Nina Moorman and Gombolay, 2023], and more. There is a groundswell of initiatives to push interactive robot learning to its limits, such as by curating large-scale datasets and leveraging foundation models.

Yet, despite decades of research, interactive robot learning systems are not deployed in the real world and fall short in practice because, among many factors, researchers typically assume humans are 'oracles' that are devoid of the myriad of human cognitive biases that confound our models. For example, researchers have shown that there is a disconnect between what people *say* they are doing and what people actually *do* when accomplishing a task [Bilalić *et al.*, 2008]. Further, humans do not understand robots – how robots perceive, make decisions, and act – known as the *correspondence problem.* Thus, interactive robot learning frameworks fail to be truly interactive as the robot cannot provide insight as an actionable feedback signal to the human.

This paper presents computational techniques and human factors insights toward democratizing robot learning: plac-

ing the power of robot learning in the hands of non-roboticist end-users. The contributions presented support the design of robots that learn from end-user interaction – without the need for a roboticist – to align the robot's behaviors and values. This paper defines the problem of aligning behaviors as synthesizing a robot control policy that produces actions, conditioned on observations of the world, that abide by what the human would want the robot to do given that information. Conversely, aligning robot values enables the robot to infer the correct constrained optimization describing the humans' desired outcomes. Behavior alignment is means-driven whereas value alignment is outcome-driven. Both are needed to enable true alignment.

The proposed framework enables robots to (1) learn models of human values and behaviors, such as through novel Inverse RL (IRL) techniques, (2) explain to the human what the robot has learned through advances in interpretable and explainable Artificial Intelligence (xAI), and (3) scale this closed-loop HRI up to enabling multiple, heterogeneous humans and robots to coordinate activities at scale for ad hoc human-robot teaming. Key to this work is exploring various latent and observable (e.g., the mode of human instruction) factors that confound the ability to precisely infer and align a human end-users desired values and behaviors with a robots'. This paper argues that researchers cannot assume a *spherical human*[1] and demonstrates the tremendous advances in robot capabilities when the correspondence problem is appropriately tackled.

2 Learning from Humans to Robots

Robot learning from human interaction can serve three key supporting roles, enabling robots to acquire skills for task performance (e.g., LfD to perform 3D tasks) [Gombolay *et al.*, 2018a], anticipate future human behavior so that the robot can perform complementary actions supporting the humans [Paleja *et al.*, 2024], and even assess human task performance, e.g. for tutoring [Gombolay *et al.*, 2017] or even robots teaching humans to be better robot teachers [Schrum *et al.*, 2022a].

These aspects of robot learning can be grounded in a Markov decision process (MDP) formalism for human-robot behavior and value alignment. This problem of alignment takes as input an MDP without a known reward function (MDP\R) as a 4-tuple, $\langle S, A, T, \gamma, \rho \rangle$, where $s \in S$ is a state of the world, $a \in A$ is an action that can be performed, $T : S \times A \times S' \rightarrow [0, 1]$ is the probability of transitioning s' given action a in s, a discount factor, γ, prioritizing accruing short-vs. long-term reward, and $\rho : S \rightarrow [0, 1]$ is a distribution over initial states. The reward function, $R : S \times A \times S' \rightarrow \mathbb{R}$, is omitted as it is a latent variable in the human teacher's mind.

The goal for the robot learner is to take this MDP \R along with human data, $\mathcal{D}$ (e.g., human demonstrations as a sequence of human states or actions, natural language, etc.) and output either an aligned behavioral policy, $\pi^* : S \times A \rightarrow [0, 1]$, as a probability distribution over actions aligned with human expectations or an aligned value (i.e., reward) function, R^*, where $*$ denotes that the policy is perfectly aligned. There are a variety of techniques for reverse engineering a policy or reward function including inverse RL (IRL) which seeks to learn a reward function and policy simultaneously, and Behavior Cloning, which learns a policy directly.

However, reverse engineering humans' values or behavioral policies is confounded by the stochastic, inconsistent, and misleading data people provide. As noted earlier, prior work has shown people have difficulty faithfully describing their decision-making processes [Bilalić *et al.*, 2008]. For applications of interactive robot learning in expert domains (e.g., healthcare), naturalistic decision-making research shows that experts may not even think in a rules-based fashion that can be articulated [Klein, 2008]. Lastly, even experts exhibit suboptimal and diverse strategies for accomplishing the same task, which can thwart robot learning algorithms that are best suited for learning from idealized data.

This section presents state-of-the-art technical approaches to address the need to learn from suboptimal and diverse forms of human feedback (Section 2.1) as well as guidelines from human-subject experiments for how to design for successful interactive robot learning (Section 2.2).

2.1 Learning from Diverse, Suboptimal Humans

Prior work has shown that even human experts, e.g. commercial aviation pilots, exhibit behavior so diverse on the same task that it is more practical to learn separate robot policies aligned to each human rather than pooling data across demonstrators [Sammut *et al.*, 1992]. However, this is far from ideal, as end-users would then need to train each robot from scratch, which may be time- and cost-prohibitive. Ideally, interactive robot learning methods should be able to leverage data from a cohort of demonstrators to mitigate the curse of dimensionality and achieve a higher degree of personalized alignment.

Towards this vision of sample-efficient, personalized alignment with diverse demonstrators, this paper presents Multi-Style Reward Distillation (MSRD) [Chen *et al.*, 2020], which learns from a dataset of humans performing a common task but doing so while exhibiting varying preferences or strategies. The key to MSRD is employing neural network distillation in an adversarial IRL [Fu *et al.*, 2018] setup in which a reward function-based discriminator learns from all demonstrators what is *common* (i.e., the task reward) while preference-specific discriminators learn the components of the humans' rewards that are strategy-specific (i.e., the strategy reward). MSRD was extended to relax the assumption of annotated strategies and allow for online, incremental learning from a growing population of end-users with Fast Lifelong Adaptive Inverse Reinforcement learning from demonstrations (FLAIR) [Chen *et al.*, 2023]. FLAIR demonstrated on a physical robot table tennis setup for teaching strokes that FLAIR objectively and subjectively performs human-robot alignment better than user-specific models.

Suboptimal performance in teaching robots is a second challenge for robot learning. An effective strategy in cases where human suboptimality is due to problem scale or workload [Molina *et al.*, 2018] is to have robots learn a size-invariant behavior model from human task performance on

[1]Derived from the physics joke: "Assume a spherical cow..."

smaller problems and scale that learned model up to larger problem sizes. This is an effective strategy when human workload is a key driver of suboptimality and has demonstrated effectiveness for human-robot alignment in healthcare [Gombolay *et al.*, 2018b], manufacturing [Gombolay *et al.*, 2018a], and military domains [Gombolay *et al.*, 2018a].

When suboptimality is intrinsic to the human's demonstration of the task, regardless of workload, Self-Supervised Reward Regression (SSRR) is an effective LfD technique for learning to perform a task better than what the human was able to demonstrate without needing to ask the human what "better" means [Chen *et al.*, 2021]. Inspired by prior work by [Brown *et al.*, 2019], SSRR learns an idealized reward function by imitating the human demonstrator, automatically generating worse behavior by adding noise to the robot's actions, and then inferring what the behavior would look like if it were less noisy than the human demonstration. The key to SSRR is a proper characterization of the noise-performance relationship and accounting for the covariate shift induced in learning from real human and synthetic, noisy data to extrapolate beyond the performance of the human demonstrator.

Robots can also account for suboptimality by correcting the human data before learning the behavior. Mutual INformation-Driven MEta-Learning from Demonstration (MIND MELD) is a technique in which robots learn a personalized calibration model of corrective human actions (e.g., as delayed/anticipatory or over-/under-corrective) and then apply that model to compensate for each person's unique suboptimality on a novel LfD task [Schrum *et al.*, 2022b].

Lastly, safety is critical for interactive robot learning with real end-users. Offline learning, in which the robot does not need to learn through trial and error, is a valuable approach in safety-critical applications of machine learning (ML). Dual Reward and policy Offline Inverse Distillation (DROID) is a technique that grounds the framework of MSRD in an offline setting as was successfully applied to inferring the personalized values of NASA Jet Propulsion Laboratory Rover Planners in how they design paths for the Mars Curiosity Rover [Jayanthi *et al.*, 2023]. Safe behavior and values can also be taught by human demonstrators, such as by jointly inferring the human demonstrator's preferences and constraints for a task in the form of a reward function and a set of neural certificate barrier functions in a framework called SECURE (ShiElding with Control barrier fUnctions in inverse REinforcement learning) [Yang *et al.*, 2024].

2.2 Human-centered Design Principles

It is critical for researchers to study non-roboticists and characterize the psychology of interactive robot learning. Motivated by a focus on developing assistive robots for ADL, researchers have worked with older and younger adults to ascertain their attitudes towards what modes of robot learning interactions are most usable and how those modes impact perceptions of robot anthropomorphism and user workload [Moorman *et al.*, 2023; Botti *et al.*, 2024]. Importantly, blame attribution depends on demographics: some populations are more likely to blame themselves for a robot's failure than to blame the robot (or the engineer who developed the robot system). Further, people bestow upon the robot a sense of agency in which a person judges their own competence as a teacher based upon what they believe the robot thinks about them [Hedlund *et al.*, 2021]. These experiments show the importance of applying universal design principles to account for homophily and the role of end-user personality traits and backgrounds to achieve democratized robot learning [Schrum *et al.*, 2024; Hedlund *et al.*, 2021; Schrum *et al.*, 2021; Tambwekar and Gombolay, 2023].

3 Learning from Robots to Humans

Humans must be supported in situated learning interactions with robots. Yet, modern learning frameworks are typically one-way (i.e., human-to-robot) and black-box (i.e., humans cannot inspect the inner-workings of the robot's learning process). This section presents methods to ameliorate these limitations to achieve closed-loop interactive robot learning.

3.1 Algorithmic Insights to Support Humans

Interpretability in learning is key for safety-critical domains, where humans must inspect and simulate the inner workings of robots. DTs are a gold standard in such domains but are difficult to leverage due to a conflict between their non-differentiable structure and modern reliance on gradient-based learning [Rudin *et al.*, 2022]. New optimization techniques and models for *differentiable* DTs have been developed that learn interpretable robot behavior policies through RL [Silva *et al.*, 2020; Paleja *et al.*, 2022]. This work has been extended to Personalized Neural Trees (PNTs) for learning from diverse human demonstrations in which the DT learns and operates over state variables and personalized embeddings via variational inference [Paleja *et al.*, 2020]. These DTs can support explanatory debugging in interactive robot learning [Tambwekar and Gombolay, 2023].

As a complementary approach to model interpretability, robots can also provide explainable feedback to humans on how to improve their demonstrations to robots in LfD. Reciprocal MIND MELD leverages the MIND MELD framework to dynamically predict how a user is suboptimal on a sequence of teaching tasks and, at each iteration of the sequence, provide adaptive feedback to further enhance the quality of the human's demonstrations. This approach can improve the performance of the robot in aligning its behavior to human expectations through this iterative teaching approach [Schrum *et al.*, 2022a].

3.2 Human-centered Design Principles

Central to interactive robot learning is the need for the human teacher to develop a mental model of the robot's learning process. This mental model supports the human in debugging misalignment in robot values or behavior. Yet, prior work has shown that humans can get *worse* at teaching robots when trouble-shooting only based upon watching the robots' attempts at performing tasks with black-box models [Gopalan *et al.*, 2022]. As such, it is critical for robot learning to support explanatory debugging with interpretable models (e.g., DTs) or other faithful, xAI techniques.

A challenge here is that explainability is in the eyes of the beholder [Silva *et al.*, 2023]. Various types of explainabil-

ity, such as decision trees (DTs) and counterfactual reasoning, can have a positive or *negative* impact on end-users depending on demographics [Gombolay *et al.*, 2024]. Users can even exhibit counterproductive preferences for how a robot explains its reasoning [Silva *et al.*, 2024]. Thus, it is critical to take a human-centered approach to developing xAI.

4 Scaling from Dyads to Teams

To scale interactive and interpretable robot learning for teaming, e.g. for disaster response [Seraj *et al.*, 2023], manufacturing [Gombolay *et al.*, 2018a], and healthcare [Gombolay *et al.*, 2018b], these frameworks must account for "non-spherical" aspects of human task performance and support humans in communicating their values for team coordination.

4.1 Modeling Human Performance in Teams

The performance characteristics of human work in industrial settings are stochastic and time-varying. Season manufacturing is a prime example where workers must be hired, trained, and employed over months during which their performance never plateaus. As such, robots that learn to coordinate the activities of human-robot teams in these settings ought to not only model this variable performance but *explore* which workers are best suited for which tasks. Prior work has developed models and team scheduling frameworks that perform this integrated modeling of human performance and an adaptive, extended Kalman filter to capture that humans are stochastic and learn (and generalize that knowledge) to perform tasks better with repetition [Liu *et al.*, 2021]. In validating this approach, the researchers found that human subjects working with robot teammates in an analog manufacturing setup favored robots that effectively balance this exploration-vs.-exploitation trade-off [Liu *et al.*, 2021].

This work was recently extended to leverage graph neural networks (GNNs) for function approximation in learning to coordinate these human-robot teams [Altundas *et al.*, 2022]. GNNs are apt in learning coordination policies via RL [Altundas *et al.*, 2022] or from demonstration [Wang *et al.*, 2022] on smaller-scale coordination problems and scaling that learned model to large-scale problems – a technique for handling suboptimality discussed in Section 2.1. Prior work has also shown that GNNs can be parameterized with the diverse values of human operators for alignment of team coordination [Wang *et al.*, 2022].

4.2 Supporting Humans in Aligning Coordination

While interactive robot learning techniques are typically applied for robot skill learning with a single human teaching a single robot, these approaches can be designed to learn from humans how to coordinate teams of robots. As presented in Section 2.1, PNTs were designed to learn from human scheduler demonstrations for how those humans would coordinate mixed teams of human and robot workers [Paleja *et al.*, 2020]. PNTs are also interpretable once optimized into a discrete, symbolic, tree structure, which supports humans in explanatory debugging. Users could explicitly modify the tree through a user interface or replace/augment demonstrations that did not result in the intended learning effect.

Moving away from typical, symbolic scheduling tasks, researchers have also sought to develop interactive techniques for humans to program distributed teams operating under partial observability and communication limitations (e.g., due to limited radio and sensor ranges) [Seraj *et al.*, 2023]. In such domains, teaming requires the distributed team members to act both physically within the environment and communicate with their team members to develop team situational awareness and properly coordinate activities. However, it is difficult for humans to teach their values for acting and communicating simultaneously, as demonstrated by prior work [Seraj *et al.*, 2023]. This work developed a remedy: Mixed-iniTiative mUlti-agent appRenticeship lEarning (MixTURE), which supports the human in aligning the values of a distributed, multi-agent team. The approach works by affording humans global situational awareness and a programming interface to command robot actions. These robots then must learn to align their behavior with the human's instructions by (1) automatically learning how to communicate the right information and to whom and (2) learning *decentralized* action and communication models, as there is no global situational awareness at test time. MixTURE showed positive subjective and objective results against baselines in application to a virtual, multi-agent wildfire fighting environment.

5 Discussion, Limitations, and Future Work

This paper has presented the need for and approaches for achieving democratized robot learning through interactivity and interpretability. Critical to the success of this work is taking a human-centered approach and conducting human-subject experiments with non-roboticists to inform the design of and validate these interactive and interpretable robots.

Despite the advances presented in this paper, interactive and interpretable robot learning is still hiding within research laboratories rather than actualized in homes and workplaces. This limitation is due to a numerous list of challenges, including the cost of robot hardware, the need for even better sample efficiency, the difficulty in collecting real-world data, and the fact that robots do not have the inductive biases and common-sense reasoning abilities that are part and parcel with human intelligence. In future work, research in interactive robot learning needs to suss out whether foundation models are the proverbial holy grail as a foundation for robot learning or merely a detour. Research in interpretable ML is still in its infancy and, while DTs are powerful, ultimately it is difficult to precisely explain complex behavior in a sparse model. Thus, further work is needed to develop the right levels of abstraction for the right context to support explanatory debugging and appropriate trust in robotic systems.

There are also ethical issues that must be addressed to safely deploy robotic systems. These issues include the need to identify [Hundt *et al.*, 2022] and remove [Silva *et al.*, 2022] toxic reasoning in foundation models governing robot behavior. Further, robots supplanting human work could result in psychological and material harm [Schrum *et al.*, 2021]. These and other ethical issues must be judiciously studied to afford a safe rollout of these emerging technologies.

References

[Altundas *et al.*, 2022] Batuhan Altundas, Zheyuan Wang, Joshua Bishop, and Matthew Gombolay. Learning coordination policies over heterogeneous graphs for human-robot teams via recurrent neural schedule propagation. In *Proc. Int'l Conf. Intelligent Robots and Systems*, pages 11679–11686, 2022.

[Bilalić *et al.*, 2008] Merim Bilalić, Peter McLeod, and Fernand Gobet. Inflexibility of experts—reality or myth? quantifying the einstellung effect in chess masters. *Cognitive psychology*, 56(2):73–102, 2008.

[Botti *et al.*, 2024] Erin Botti, Lakshmi Seelam, Chuxuan Yang, Nathaniel Belles, Zulfiqar Zaidi, and Matthew Gombolay. Assistive robot learning from demonstrations by older adults: Feasibility in younger and older cohorts. In *Robotics: Science and Systems*, 2024.

[Brown *et al.*, 2019] Daniel Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *Int'l Conf. ML*, 2019.

[Chen *et al.*, 2020] Letian Chen, Rohan Paleja, Muyleng Ghuy, and Matthew Gombolay. Joint goal and strategy inference across heterogeneous demonstrators via reward network distillation. In *ACM/IEEE Int'l Conf. HRI*, pages 659–668, 2020.

[Chen *et al.*, 2021] Letian Chen, Rohan Paleja, and Matthew Gombolay. Learning from suboptimal demonstration via self-supervised reward regression. In *Conf. Robot Learning*, pages 1262–1277. PMLR, 2021.

[Chen *et al.*, 2023] Letian Chen, Sravan Jayanthi, Rohan R Paleja, Daniel Martin, Viacheslav Zakharov, and Matthew Gombolay. Fast lifelong adaptive inverse reinforcement learning from demonstrations. In *Conf. Robot Learning*, pages 2083–2094, 2023.

[Fu *et al.*, 2018] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adverserial inverse reinforcement learning. In *Int'l Conf. Learning Representations*, 2018.

[Gombolay *et al.*, 2017] Matthew Gombolay, Reed Jensen, Jessica Stigile, Sung-Hyun Son, and Julie Shah. Learning to tutor from expert demonstrators via apprenticeship scheduling. In *Workshops at the AAAI Conf. AI*, 2017.

[Gombolay *et al.*, 2018a] Matthew Gombolay, Reed Jensen, Jessica Stigile, Toni Golen, Neel Shah, Sung-Hyun Son, and Julie Shah. Human-machine collaborative optimization via apprenticeship scheduling. *Journal of Artificial Intelligence Research*, 63:1–49, 2018.

[Gombolay *et al.*, 2018b] Matthew Gombolay, Xi Jessie Yang, Bradley Hayes, Nicole Seo, Zixi Liu, Samir Wadhwania, Tania Yu, Neel Shah, Toni Golen, and Julie Shah. Robotic assistance in the coordination of patient care. *Int'l Journal of Robotics Research*, 37(10):1300–1316, 2018.

[Gombolay *et al.*, 2024] Grace Y Gombolay, Andrew Silva, Mariah Schrum, Nakul Gopalan, Jamika Hallman-Cooper, Monideep Dutt, and Matthew Gombolay. Effects of explainable artificial intelligence in neurology decision support. *Annals of Clinical and Translational Neurology*, 11(5):1224–1235, 2024.

[Gopalan *et al.*, 2022] Nakul Gopalan, Nina Moorman, Manisha Natarajan, and Matthew Gombolay. Negative result for learning from demonstration: Challenges for end-users teaching robots with task and motion planning abstractions. In *Robotics: Science and Systems*, 2022.

[Hedlund *et al.*, 2021] Erin Hedlund, Michael Johnson, and Matthew human subjectGombolay. The effects of a robot's performance on human teachers for learning from demonstration tasks. In *ACM/IEEE Int'l Conf. HRI*, pages 207–215, 2021.

[Hundt *et al.*, 2022] Andrew Hundt, William Agnew, Vicky Zeng, Severin Kacianka, and Matthew Gombolay. Robots enact malignant stereotypes. In *ACM Conf. Fairness, Accountability, and Transparency*, pages 743–756, 2022.

[Institute of Medicine, 2010] Institute of Medicine. The future of nursing: Leading change, advancing health. Nat'l Acad. Sciences, Engineering, and Medicine. 2010.

[Jayanthi *et al.*, 2023] Sravan Jayanthi, Letian Chen, Nadya Balabanska, Van Duong, Erik Scarlatescu, Ezra Ameperosa, Zulfiqar Haider Zaidi, Daniel Martin, Taylor Keith Del Matto, Masahiro Ono, and Matthew Gombolay. Droid: Learning from offline heterogeneous demonstrations via reward-policy distillation. In *Conf. Robot Learning*, pages 1547–1571, 2023.

[Klein, 2008] Gary Klein. Naturalistic decision making. *Human factors*, 50(3):456–460, 2008.

[Liu *et al.*, 2021] Ruisen Liu, Manisha Natarajan, and Matthew C Gombolay. Coordinating human-robot teams with dynamic and stochastic task proficiencies. *ACM Trans. on HRI*, 11(1):1–42, 2021.

[Molina *et al.*, 2018] Rose Molina, Matthew Gombolay, Jennifer Jonas, Julie Shah, Toni Golen, and Neel Shah. Association between labor and delivery unit census and delays in patient management Findings from a computer simulation module. *OBGYN*, 131(3):545–552, 2018.

[Moorman *et al.*, 2023] Nina Moorman, Erin Hedlund-Botti, Mariah Schrum, Manisha Natarajan, and Matthew C Gombolay. Impacts of robot learning on user attitude and behavior. In *ACM/IEEE Int'l Conf. HRI*, pages 534–543, 2023.

[Nina Moorman and Gombolay, 2023] Aman Singh Erin Hedlund-Botti Mariah Schrum Chuxuan Yang Lakshmi Seelam Nina Moorman, Nakul Gopalan and Matthew Gombolay. Investigating the impact of experience on a user's ability to perform hierarchical abstraction. In *Robotics: Science and Systems*, 2023.

[Paleja *et al.*, 2020] Rohan Paleja, Andrew Silva, Letian Chen, and Matthew Gombolay. Interpretable and personalized apprenticeship scheduling: Learning interpretable scheduling policies from heterogeneous user demonstrations. In *Advances in Neural Information Processing Systems*, volume 33, pages 6417–6428, 2020.

[Paleja *et al.*, 2022] Rohan Paleja, Yaru Niu, Andrew Silva, Chace Ritchie, Sugju Choi, and Matthew Gombolay. Learning interpretable, high-performing policies for autonomous driving. In *Robotics: Science and Systems*, 2022.

[Paleja *et al.*, 2024] Rohan Paleja, Michael Munje, Kimberlee Chang, Reed Jensen, and Matthew Gombolay. Designs for enabling collaboration in human-machine teaming via interactive and explainable systems. *arXiv:2406.05003*, 2024.

[Prince and Fantom, 2014] William Prince and Neil Fantom. World development indicators. the world bank, 2014.

[PRNewswire, 2016] PRNewswire. Genworth 2016 cost of care survey. https://prn.to/4cAwPfJ, 2016. Visited on June, 24^{th}, 2024.

[Rudin *et al.*, 2022] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16:1–85, 2022.

[Sammut *et al.*, 1992] Claude Sammut, Scott Hurst, Dana Kedzier, and Donald Michie. Learning to fly. In *ML Proc.*, pages 385–393. Elsevier, 1992.

[Schrum *et al.*, 2021] Mariah L Schrum, Glen Neville, Michael Johnson, Nina Moorman, Rohan Paleja, Karen M Feigh, and Matthew C Gombolay. Effects of social factors and team dynamics on adoption of collaborative robot autonomy. In *ACM/IEEE Int'l Conf. HRI*, pages 149–157, 2021.

[Schrum *et al.*, 2022a] Mariah L Schrum, Erin Hedlund-Botti, and Matthew Gombolay. Reciprocal MIND MELD: Improving learning from demonstration via personalized, reciprocal teaching. In *Conf. Robot Learning*, pages 956–966, 2022.

[Schrum *et al.*, 2022b] Mariah L Schrum, Erin Hedlund-Botti, Nina Moorman, and Matthew C Gombolay. Mind meld: Personalized meta-learning for robot-centric imitation learning. In *ACM/IEEE Int'l Conf. HRI*, pages 157–165, 2022.

[Schrum *et al.*, 2024] Mariah L Schrum, Emily Sumner, Matthew C Gombolay, and Andrew Best. Maveric: A data-driven approach to personalized autonomous driving. *IEEE Trans. on Robotics*, 2024.

[Seraj *et al.*, 2023] Esmaeil Seraj, Jerry Xiong, Mariah Schrum, and Matthew Gombolay. Mixed-initiative multi-agent apprenticeship learning for human training of robot teams. In *Advances in Neural Information Processing Systems*, volume 36, pages 35426–35440, 2023.

[Seraj *et al.*, 2024] Esameil Seraj, Kin Man Lee, Zulfiqar Zaidi, Qingyu Xiao, Zhaoxin Li, Arthur Nascimento, Sanne van Waveren, Pradyumna Tambwekar, Rohan Paleja, Devleena Das, and Matthew Gombolay. Interactive and explainable robot learning: A comprehensive review. *Foundations and Trends® in Robotics*, 12(2-3):75–349, 2024.

[Silva *et al.*, 2020] Andrew Silva, Taylor Killian, Ivan Jimenez, Sung-Hyun Son, and Matthew Gombolay. Optimization methods for interpretable differentiable decision trees applied to reinforcement learning. In *Int'l Conf. AI and Statistics*, pages 1855–1865, 2020.

[Silva *et al.*, 2021] Andrew Silva, Nina Moorman, William Silva, Zulfiqar Zaidi, Nakul Gopalan, and Matthew Gombolay. Lancon-learn: Learning with language to enable generalization in multi-task manipulation. *IEEE Robotics and Automation Letters*, 7(2):1635–1642, 2021.

[Silva *et al.*, 2022] Andrew Silva, Rohit Chopra, and Matthew Gombolay. Cross-loss influence functions to explain deep network representations. In *Int'l Conf. AI and Statistics*, pages 1–17, 2022.

[Silva *et al.*, 2023] Andrew Silva, Mariah Schrum, Erin Hedlund-Botti, Nakul Gopalan, and Matthew Gombolay. Explainable artificial intelligence: Evaluating the objective and subjective impacts of xai on human-agent interaction. *Int'l Journal of Human–Computer Interaction*, 39(7):1390–1404, 2023.

[Silva *et al.*, 2024] Andrew Silva, Pradyumna Tambwekar, Mariah Schrum, and Matthew Gombolay. Towards balancing preference and performance through adaptive personalized explainability. In *ACM/IEEE Int'l Conf. HRI*, pages 658–668, 2024.

[Skakkebæk *et al.*, 2022] Niels E Skakkebæk, Rune Lindahl-Jacobsen, Hagai Levine, Anna-Maria Andersson, Niels Jørgensen, Katharina M Main, Øjvind Lidegaard, Laerke Priskorn, Stine A Holmboe, Elvira V Bräuner, et al. Environmental factors in declining human fertility. *Nature Reviews Endocrinology*, 18(3):139–157, 2022.

[Takayama *et al.*, 2008] Leila Takayama, Wendy Ju, and Clifford Nass. Beyond dirty, dangerous and dull: what everyday people think robots should do. In *ACM/IEEE Int'l Conf. HRI*, pages 25–32, 2008.

[Tambwekar and Gombolay, 2023] Pradyumna Tambwekar and Matthew Gombolay. Towards reconciling usability and usefulness of explainable AI methodologies. *arXiv preprint arXiv:2301.05347*, 2023.

[Tambwekar *et al.*, 2023] Pradyumna Tambwekar, Andrew Silva, Nakul Gopalan, and Matthew Gombolay. Natural language specification of reinforcement learning policies through differentiable decision trees. *IEEE Robotics and Automation Letters*, 2023.

[Wang *et al.*, 2022] Zheyuan Wang, Chen Liu, and Matthew Gombolay. Heterogeneous graph attention networks for scalable multi-robot scheduling with temporospatial constraints. *Autonomous Robots*, 46(1):249–268, 2022.

[Wright, 2023] James Wright. *Robots won't save Japan: an ethnography of eldercare automation*. Cornell University Press, 2023.

[Yang *et al.*, 2024] Yue Yang, Letian Chen, Zulfiqar Zaidi, Sanne van Waveren, Arjun Krishna, and Matthew Gombolay. Enhancing safety in learning from demonstration algorithms via control barrier function shielding. In *ACM/IEEE Int'l Conf. HRI*, pages 820–829, 2024.