Appendix

Mariah L. Schrum, Erin Hedlund-Botti, Matthew C. Gomoblay
Institute for Robotics and Intelligent Machines
Georgia Institute of Technology
mschrum3@gatech.edu, ehedlund6@gatech.edu, matthew.gombolay@cc.gatech.edu

1 Related Work

Prior work has shown that robot-centric learning from demonstration (LfD) outperforms human-centric LfD in terms of the number of mistakes an agent makes when demonstrations are high-quality [1]. However, Laskey et al. [2] illustrated that human demonstrators tend to provide low quality corrective demonstrations to a robot which produces poor learning results for robot-centric LfD approaches, such as DAgger [1]. Prior work has investigated methods for improving upon a teacher's ability to provide high-quality demonstrations. For example, Spencer et al. [3], Laskey et al. [4], and Kelly et al. [5] developed methods for robot-centric LfD that are more human-friendly and easier for the demonstrator to understand, thereby improving the quality of the demonstrations. Additionally, Spencer et al. [3] and Grollman and Billard [6] investigated how to gain information from failure trajectories. However, these methods do not provide the demonstrator with feedback about how to improve their demonstrations if they are suboptimal. To address this problem, Amershi et al. posited that transparency from a learning system can improve participants' opinions of the system and improve demonstrations [7]. Reciprocal MIND MELD capitalizes on the ideas of Amershi et al. to increase transparency and provide constructive feedback to the demonstrator to improve upon the quality of the demonstrations provided to the robot learner.

Other approaches have introduced methods for agents to better learn from suboptimal demonstrations via inverse-reinforcement learning (IRL). In Chen et al., the authors introduce Self-Supervised Reward Regression (SSRR) in which the authors improve upon an agent's ability to learn from suboptimal demonstrations by characterizing the relationship between noise and performance [8]. Their approach bootstraps off of suboptimal demonstrations to learn an idealized reward function. Similarly, T-Rex and D-Rex improve upon the ability to learn from suboptimal demonstrations by learning a reward function from a ranked set of demonstrations [9, 10]. Additionally, Burchfiel et al. demonstrate how to learn from ranked demonstrations where the scoring is noisy or suboptimal [11]. To incorporate preferences, Myers et al. presented an algorithm that learns a multimodal reward function [12]. Valko et al. developed a semi-supervised algorithm IRL to learn from both expert trajectories and unlabeled or suboptimal trajectories [13].

To improve upon suboptimal demonstrations in imitation learning, Kaiser and Dillmann characterize the ways in which a demonstrator can be suboptimal and presents methods to cope with this suboptimality [14]. To better learn from suboptimal demonstrators in a robot-centric paradigm, Menda et al. proposed EnsembleDAgger which approximates a Gaussian Process via an ensemble of neural networks [15]. The authors utilize the variance from this ensemble as a metric of confidence for a novice demonstrator. Additionally, Schrum et al. introduced MIND MELD [16], a personalized algorithm which meta-learns an individual-specific embedding describing a teacher's suboptimal tendencies. MIND MELD utilizes this embedding to map suboptimal labels to better labels and was able to outperform prior work when teaching an agent to perform a task in a driving simulator domain. In Reciprocal MIND MELD, we aim to aid the demonstrator in providing higher quality demonstrations instead of correcting for suboptimality under-the-hood.

Several approaches have also investigated how best to provide feedback to a demonstrator to provide better demonstrations [17, 18]. Cakmak and Takayama conducted a study investigating several modalities for communicating improvements to a demonstrator. The authors found instructional videos to be the best modality for improving demonstrators' teaching abilities [17]. Sena et al. investigated video feedback with and without rule guidance and found that both modalities produced

6th Conference on Robot Learning (CoRL 2022), Auckland, New Zealand.

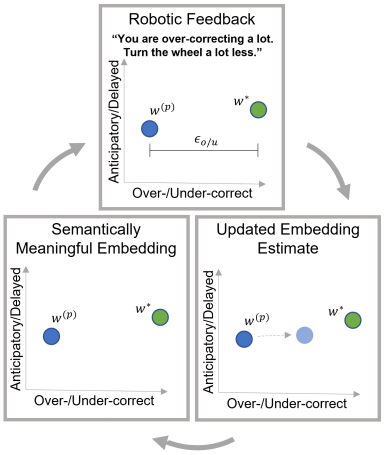


Figure 1: This figure shows our Reciprocal MIND MELD framework. $\epsilon_{o/u}$ is the distance between the participant's current embedding, $w^{(p)}$, and the perfect embedding, w^*, along the over-/under-correcting dimension.

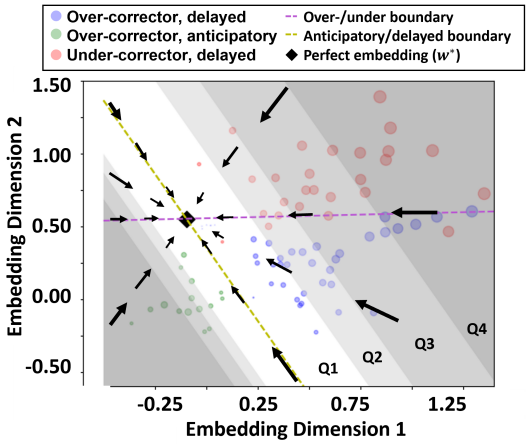


Figure 2: This figure depicts the learned embedding space and decision boundaries. Each point represents the embedding of a demonstrator, and the diameter represents the magnitude by which participants are anticipatory/delayed. Q1-Q4 indicate quartiles one through four for the anticipatory/delayed dimension.

better results than no feedback [18]. In our Reciprocal MIND MELD architecture, we employ verbal feedback. In future work, we plan to investigate how different feedback modalities impact Reciprocal MIND MELD's outcomes.

2 Driving Simulator Domain

In keeping with prior work [19], we utilize a driving simulator domain to evaluate our Reciprocal MIND MELD architecture. We utilize the high-fidelity physics simulator, Airsim with Unreal Engine and an Xbox steering wheel. In this domain, participants are tasked with teaching a car to drive from a start location to a goal in various environments while avoiding obstacles. The action space consists of the position of the wheel (-540° to 540°), and the state space consists of images from the car's first-person perspective, position, velocity, and acceleration.

3 Reciprocal MIND MELD Architecture

Fig. 1 shows the steps in the Reciprocal MIND MELD framework. To determine the robotic feedback that should be provided to the demonstrator, we first learn a semantically meaningful embedding space. The robot then provides feedback to the demonstrator based upon the distance from the perfect embedding in each semantically meaningful dimension. For example, the robot provides feedback in the over-/under-correcting dimension based on the distance, $\epsilon_{o/u}$. We then re-estimate the embedding after robotic feedback. In Study 1 and Study 2, participants experience four and five rounds of robotic feedback respectively. Between rounds, if the participant improves their feedback but is still not within the first quartile, the robot says, "That is better but..." followed by the appropriate feedback as shown in Table 1.

Table 1 shows the feedback provided to the demonstrator in Study 1 for the over-/under-correcting dimension. If a demonstrator is in a quartile that is farther from the perfect demonstrator, the feedback is intended to shift their embedding by a larger amount than demonstrators in quartiles closer to the perfect demonstrator. Analogous feedback is provided for the anticipatory/delayed dimension in Study 2 and Study 3. In all conditions, regardless of whether feedback is provided, the robot interacts with the participant and says "Please provide me with a demonstration" before each round.

Fig. 2 illustrates the embedding space in which the size of the points represents the magnitude by which a participant is anticipatory/delayed. Q1-Q4 indicate the quartiles for anticipatory/delayed.

Points that are farther from the decision boundaries represent participants who are more suboptimal in their demonstrations. The objective is to provide robotic feedback to shift a participants embedding into Q1.

Table 1: This table shows the feedback a participant receives based on their quartile and study condition for Study 1. Analogous feedback for the Cooperative condition is provided in Study 2 for the anticipatory/delayed dimension in addition to the over-/under-correcting dimension.

Cooperative Quartile	**Adversarial Quartile**	**Robotic Feedback**
First	Fourth	"Your feedback is good! Keep it up."
Second	Third	"You are slightly over-/under-correcting. Please turn the wheel a bit more/less."
Third	Second	"You are over-/under-correcting. Please turn the wheel more/less."
Fourth	First	"You are over-/under-correcting a lot. Please turn the wheel a lot more/less."

4 MIND MELD Architecture

Below we describe the MIND MELD architecture and discuss the alterations to learn a semantically meaningful embedding space.

4.1 Network Architecture

Fig. 3 shows the MIND MELD architecture. The three main components of the architecture are: 1) the bi-directional long short-term memory (LSTM) encoder, $\mathcal{E}_{\phi'}\colon A \rightarrow Z$, 2) the prediction subnetwork, $f_\theta\colon Z \times W \rightarrow \mathbb{R}$, and 3) the mutual information subnetwork, $q_\phi\colon Z \times \mathbb{R} \rightarrow \mathcal{N}_W$. Our goal is to improve upon the corrective feedback, $a_t^{(p)}$, from a demonstrator, p. The corrective feedback from the human demonstrator from $t - \Delta t : t + \Delta t$ is fed into the bi-directional LSTM, $\mathcal{E}_{\phi'}$, to extract an encoding, $z^{(p)}_{t-\Delta t:t+\Delta t}$. The f_θ subnetwork takes in the encoding, $z^{(p)}_{t-\Delta t:t+\Delta t}$, and the personalized embedding, $w^{(p)}$, and learns the predicted difference, $\hat{d}_t^{(p)}$, between the optimal label, o_t, and the human's corrective label, $a_t^{(p)}$. The q_ϕ subnetwork learns to map the difference, $\hat{d}_t^{(p)}$, and the encoding, $z^{(p)}_{t-\Delta t:t+\Delta t}$, to a posterior distribution over the person's embedding, $w^{(p)}$. We estimate an individual's learned embedding, $\hat{w}^{(p)}$, by sampling from the approximate posterior [20]. $w^{(p)}$ is initialized based upon the prior, $\hat{w}^{(p)} \sim \mathcal{N}(0, 1)$.

4.2 Loss Function for Semantic Meaning

To learn a semantically meaningful embedding space, we add an additional network head, p_ψ, to the MIND MELD architecture to aid in learning the embedding space. p_ψ is a linear layer to encourage the embedding space to be linearly separable. We utilize a mean squared error (MSE) loss, $l = \frac{1}{N}\sum_i \left(p_\psi(w^{(i)}) - m^{(i)}_{o/u,a/d}\right)^2$, to train the network to predict the suboptimal tendencies, $m_{o/u}$ and $m_{a/d}$, (i.e., the magnitude by which a demonstrator over-/under-corrects and is delayed/anticipatory) given the personalized embedding. We calculate $m_{o/u}$ and $m_{a/d}$ via dynamic time warping (DTW) [21] between the demonstrations and the optimal labels in the calibration tasks. This loss helps to ensure that our embedding space can be translated into actionable robotic feedback, (i.e., the magnitude by which a demonstrator over-/under-corrects and is delayed/anticipatory) given the personalized embedding.

4.3 Variational Inference

We assume that humans provide heterogeneous and distinct styles when providing corrective feedback to the robot. A person's corrective style is encapsulated in the embedding, $w^{(p)}$, for person, p. To learn $w^{(p)}$, we maximize the lower bound on the mutual information between the learned embedding, $w^{(p)}$, and the predicted difference between the human feedback and the optimal feedback,

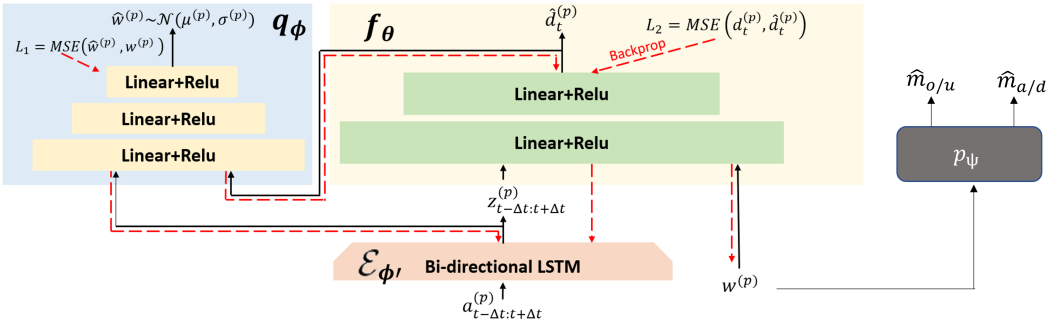


Figure 3: This figure shows the MIND MELD network architecture. The inputs to the architecture are a demonstrator, p's, corrective labels, $a^{(p)}_{(t-\Delta t:t+\Delta t)}$, from time $t-\Delta t$ to $t+\Delta t$ and the personalized embedding, $w^{(p)}$. The bi-directional LSTM extracts sequential information about the demonstrator's feedback. The f_θ subnetwork learns the predicted difference, $\hat{d}_t^{(p)}$, by minimizing the mean squared error (MSE) between $\hat{d}_t^{(p)}$ and the true difference, $d_t^{(p)} = a_t^{(p)} - o_t$, between the demonstrator's corrective feedback, $a_t^{(p)}$, and the optimal label, o_t. The re-creation subnetwork q_ϕ maximizes mutual information between the personalized embedding, $w^{(p)}$, the encoding $z^{(p)}_{(t-\Delta t:t+\Delta t)}$, and the learned difference, $\hat{d}_t^{(p)}$ to estimate the learned embedding, $\hat{w}^{(p)}$ [16, 19]. We add the additional network head, p_ψ, to learn a semantically meaningful embedding space. The outputs $\hat{m}_{o/u}$ and $\hat{m}_{a/d}$ are estimates for how much a demonstrator is over-/under-correcting and anticipatory/delayed.

$\hat{d}_t^{(p)}$ (Eq. 1). Intuitively, maximizing mutual information means that observing the difference, $\hat{d}_t^{(p)}$, will reduce uncertainty about the personalized embedding.

In Eq. 1, the mutual information between $z^{(p)}$, $\hat{d}_t^{(p)}$, and personalized embedding, $w^{(p)}$, is denoted as $I(w^{(p)}; z^{(p)}, \hat{d}_t^{(p)})$. However, maximizing the mutual information requires access to an intractable posterior distribution, $P(w^{(p)}|z^{(p)}, \hat{d}_t^{(p)})$; therefore, we employ variational inference and a lower bound on mutual information to estimate the distribution using q_ϕ [22]. The variational lower bound is $L_I(f_{\theta|w}, q_{\phi|\theta})$.

$$I(w^{(p)}; z^{(p)}, \hat{d}_t^{(p)}) = H(w^{(p)}) - H(w^{(p)}|z^{(p)}, \hat{d}_t^{(p)}) \geq \\ \mathbb{E}[log(q_\phi(w^{(p)}|z^{(p)}, \hat{d}_t^{(p)}))] + H(w^{(p)}) = L_I(f_{\theta|w}, q_{\phi|\theta}) \tag{1}$$

The MIND MELD architecture utilizes two loss functions, one to learn the personalized embedding, $w^{(p)}$, and another to learn the amount by which a person's feedback is suboptimal, $\hat{d}_t^{(p)}$, as shown in Fig. 3. For the q_ϕ subnetwork, we minimize the mean squared error between the sampled approximation of the embedding, $\hat{w}^{(p)}$, and the personalized embedding, $w^{(p)}$, which is equivalent to maximizing the log-likelihood of the posterior. The loss function for the f_θ subnetwork is the mean squared error between the predicted difference, $\hat{d}_t^{(p)}$, and the difference between the human feedback and the optimal labels, $d_t^{(p)} = a_t^{(p)} - o_t$. These two losses are summed (Eq. 2) and backpropagated through the layers and the input embedding, $w^{(p)}$, so that the embedding converges to reflect a person's feedback style. At test time, the MIND MELD network parameters θ, ϕ, and ϕ' are frozen. We then backpropagate only through $w^{(p)}$, to learn an embedding that encapsulates a participant's suboptimal style.

$$L_{(\theta,\phi,\phi',w)} = L_{1_{(\theta,\phi,\phi')}} + \lambda L_{2_{(\theta,\phi')}} \tag{2}$$

$$L_{1_{(\theta,\phi,\phi')}} = \frac{1}{K+1}\sum_{k=0}^{K} ||\hat{w}_k^{(p)} - w_k^{(p)}|| \tag{3}$$

$$L_{2_{(\theta,\phi')}} = ||d_k^{(p)} - \hat{d}_k^{(p)}|| \tag{4}$$

5 Calibration and Novel Tasks

Fig. 6 a-d shows the calibration tasks employed in the study. These tasks of pre-recorded policy rollouts and are consistent for all participants. Fig. 6 e-g shows the novel tasks in which participants provide demonstrations to teach the car to get from the start location to the goal.

6 Additional Results from Study 1

Table 2: This table shows the mean, (standard deviation), and test statistics of the subjective metrics and $\Delta\epsilon_{o/u}$ for Study 1. Δ Trust and Δ Fluency describe the change in Trust and Fluency respectively between rounds one and four.

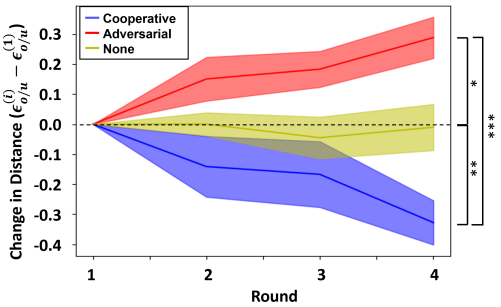
	Cooperative	**Adversarial**	**None**	Test Statistic	p-value
$\boldsymbol{\Delta\epsilon_{o/u}}$	0.33 (0.2)	-0.30 (0.2)	0.01 (0.2)	$F(2,24) = 20.2$	$\boldsymbol{p < .001}$
Workload	37.5 (16.4)	46.1 (19.5)	53.5 (11.6)	$F(2,24) = 2.21$	$p = .132$
Likeability	6.69 (2.0)	6.81 (1.5)	6.86 (1.4)	$F(2,24) = .024$	$p = .978$
Intelligence	6.31 (1.6)	5.57 (1.1)	6.24 (1.4)	$F(2,24) = 1.03$	$p = .372$
Δ **Trust**	0.56 (0.4)	-0.01 (0.4)	0.05 (0.2)	$F(2,24) = 5.15$	$\boldsymbol{p = .014}$
Δ **Fluency**	0.34 (0.4)	-0.13 (0.3)	-0.04 (0.4)	$F(2,24) = 5.10$	$\boldsymbol{p = .014}$

Fig. 4a shows the change in the distance ($\epsilon_{o/u}^{(4)} - \epsilon_{o/u}^{(1)}$) in the over-/under-correcting dimension between round one and rounds one through four. Fig. 4b shows the change between rounds one and four in the amount by which the participant over-/under-corrects as calculated via dynamic time warping (DTW) between the participant demonstrations and the optimal labels. The similarity in trends between Fig. 4a and 4b suggests that robotic feedback is not only able to shift a participant's embedding but robotic feedback is also able to alter the amount by which a participant over-/under-corrects. This finding lends support to the idea that the distance from the embedding to the decision boundary is a good measure of how much a participant over-/under-corrects.

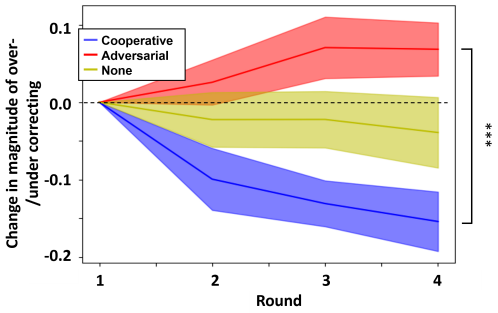
Because data does not meet parametric assumptions, we apply Friedman's test to determine if there is a statistically significant difference in how much an individual over-/under-corrects in round one versus round four as determined via DTW. We find that participants over-/under-correct significantly less in round four compared to round one in the Cooperative condition ($\chi^2(1) = 9.00, p = .003$). We find that the opposite is true in the Adversarial condition, with participants over-/under-correcting more in round four versus round one ($\chi^2(1) = 5.44$, $p = .020$). We do not find a significant difference for the None condition ($\chi^2(1) = .111, p = .740$).

We additionally compare the DTW results in round four between conditions. We find significance in an omnibus ANOVA test ($F(2,24) = 8.99$, $p = .001$). We applied Tukey post-hoc test and find that the Cooperative agent results in the participant significantly over-/under-correcting less compared to the Adversarial condition ($p < .001$). These results suggest that a participant provides demonstrations closer to the optimal by the fourth round in the Cooperative agent condition.

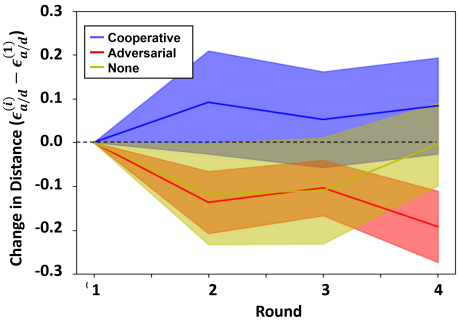
Fig. 4c and 4d show the change in the amount by which a participant provides anticipatory/delayed feedback as calculated by the distance from the perfect demonstrator in embedding space and DTW respectively. We show in Fig. 4c that as participants improve in the over-/under-correcting dimension, they tend to become worse in the anticipatory/delayed dimension when no feedback is provided. This suggests that the task of improving participants demonstration quality in both the over-/under-correcting dimension and the anticipatory/delayed dimension may be particularly diffi-



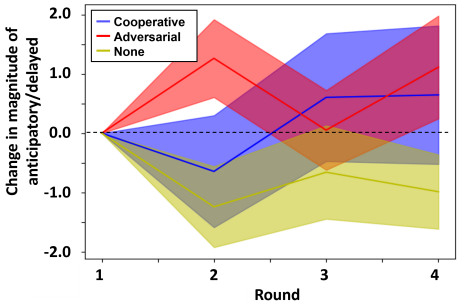
(a) Difference between the embedding distance at each round, $\epsilon_{o/u}$, and the embedding distance at round one, $\epsilon^{(1)}_{o/u}$.



(b) Amount by which participants over-/under-correct at each round minus the amount by which a participant over-/under-corrects in round one as calculated by dynamic time warping.



(c) Difference between the embedding distance at each round, $\epsilon_{a/d}$, and the embedding distance at round one, $\epsilon^{(1)}_{a/d}$.



(d) Amount by which participants are anticipatory/delayed at each round minus the amount by which a participant is anticipatory/delayed in round one as calculated by dynamic time warping.

Figure 4: This figure shows the average distance of the embedding and the dynamic time warping results for the over-/under-correcting dimension and anticipatory/delayed dimension for Study 1.

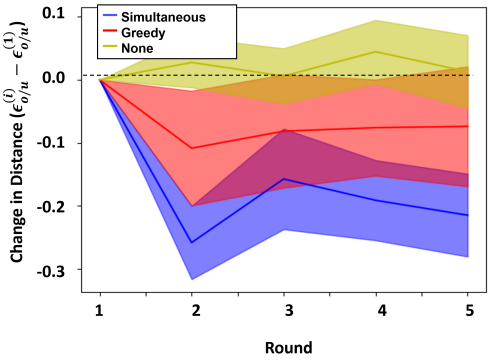
cult since improving in the over-/under-correcting dimension tends to produce greater suboptimality in the anticipatory/delayed dimension.

Table 2 shows the results of the subjective metrics. After each round, participants completed surveys measuring trust [23] and team fluency [24]. At the end of the study, participants completed surveys measuring workload [25] and likeability and perceived intelligence [26]. By applying a one-way ANOVA with Tukey post-hoc, we find that participants' trust increased significantly more ($F(2, 24) = 5.15$, $p = .014$) in Cooperative compared to Adversarial ($p = .020$) and None ($p = .038$). We do not find significance between Adversarial and None. Similar trends emerge for change in team fluency. We find that participants report statistically significantly greater positive change in fluency ($F(2, 24) = 5.10$, $p = .014$) in Cooperative compared to Adversarial ($p = .017$) and close to significant change compared to None ($p = .052$). Again, we do not find significant difference between Adversarial and None.

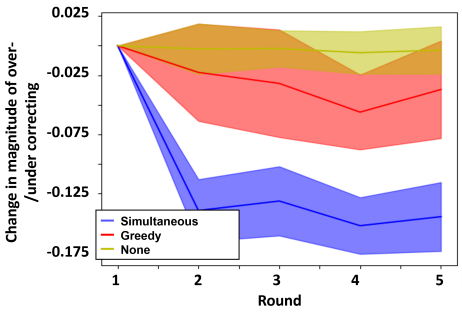
While we do not find significance between conditions with regards to the other subjective metrics, we do note some trends that merit discussion. Surprisingly, we find that Cooperative is rated as requiring lower workload compared to Adversarial and None, despite participants likely having to exert similar or additional mental effort to comply with the demands of the robot. We also find that the Cooperative robot is rated as more intelligent compared to both the Adversarial and None teachers.

7 Additional Results from Study 2

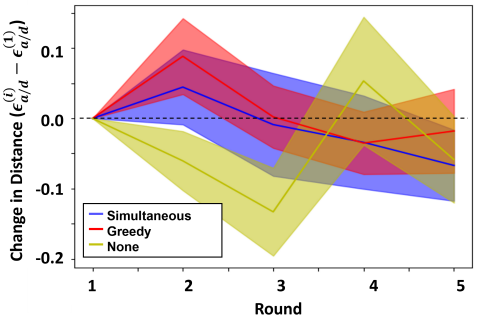
Fig. 5 shows the embedding distance in the over-/under-correcting dimension (Fig. 5a) and the anticipatory/delayed dimension (Fig. 5c). We additionally show the results of DTW for over-/under-



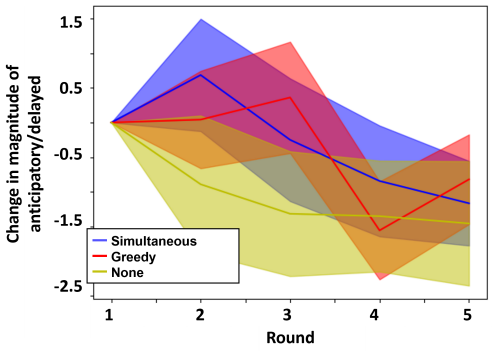
(a) Difference between the embedding distance at each round, $\epsilon_{o/u}$, and the embedding distance at round one, $\epsilon_{o/u}^{(1)}$.



(b) Amount by which participants over-/under-correct at each round minus the amount by which a participant over-/under corrects in round one as calculated by dynamic time warping.



(c) Difference between the embedding distance at each round, $\epsilon_{a/d}$, and the embedding distance at round one, $\epsilon_{a/d}^{(1)}$.



(d) Amount by which participants are anticipatory/delayed at each round minus the amount by which a participant is anticipatory/delayed in round one as calculated by dynamic time warping.

Figure 5: This figure shows the average distance of the embedding and the dynamic time warping results for the over-/under-correcting dimension and anticipatory/delayed dimension for Study 2.

correcting (Fig. 5b) and anticipatory/delayed (Fig. 5d). We find that the embedding distance in the over-/under-correcting dimension as well as the amount that a participant over-/under-corrects as determined via DTW both decrease the most in the Simultaneous condition. We also find that participants improve in the anticipatory/delayed dimension the most in the Simultaneous condition compared to Greedy and None. Although the difference between Simultaneous and None is small, this finding is noteworthy because we found in Study 1 that participants tend to become considerably worse in the anticipatory/delayed dimension as they improve in the over-/under-correcting dimension. In this study, we show that with Simultaneous feedback, participants improve in both dimensions.

Table 3 shows the mean and standard deviations of the change in the embedding distance as well as subjective metrics for each condition. Study 2 uses the same subjective metrics as Study 1: trust and team fluency after each round and workload, likeability, and perceived intelligence at the end of the study. In Study 2, we also utilize the Robot Self-Efficacy Scale to measure a participant's level of understanding [27] after each round. For each metric, we employ an ANOVA comparing the three conditions: Simultaneous, Greedy, and None. If there is a significant main effect, then we conduct a Tukey post-hoc test. As stated in the main paper, we find that Simultaneous results in significantly increased trust ($p = .032$) and team fluency ($p = .002$) ratings. Although not significant, we find that participant's understanding of the robot increased more in the Simultaneous condition compared to None and Greedy. Also, participants perceived the Simultaneous feedback as more intelligent than Greedy or None.

Table 3: This table shows the mean, (standard deviation), and test statistics of the subjective metrics, $\Delta\epsilon_{o/u+a/d}$, $\Delta\epsilon_{o/u}$, and $\Delta\epsilon_{a/d}$ for Study 2. Δ Trust, Δ Fluency, and Δ Understanding describe the change in Trust, Fluency, and Understanding respectively between rounds one and five.

	Simultaneous	**Greedy**	**None**	F(2,36)	p-value
$\boldsymbol{\Delta\epsilon_{o/u+a/d}}$	0.33 (0.25)	0.11 (0.31)	0.04 (0.28)	3.77	$\boldsymbol{p = .033}$
$\boldsymbol{\Delta\epsilon_{o/u}}$	0.267 (0.30)	0.09 (0.44)	-0.02 (0.27)	2.19	$p = .126$
$\boldsymbol{\Delta\epsilon_{a/d}}$	0.07 (0.19)	0.06 (0.23)	0.02 (0.22)	.192	$p = .826$
Workload	50.9 (12.7)	51.3 (13.7)	43.9 (17.5)	1.05	$p = .360$
Likeability	6.88 (2.16)	7.5 (1.87)	6.58 (1.66)	.790	$p = .462$
Intelligence	7.34 (2.03)	6.75 (1.72)	5.71 (1.22)	3.10	$p = .057$
Δ **Trust**	0.54 (0.60)	0.37 (0.68)	-0.77 (0.46)	3.81	$\boldsymbol{p = .032}$
Δ **Fluency**	0.78 (0.90)	0.25 (0.58)	-0.23 (0.47)	7.23	$\boldsymbol{p = .002}$
Δ **Understanding**	0.61 (0.62)	0.15 (0.58)	0.14 (0.69)	2.33	$p = .112$

8 Additional Results from Study 3

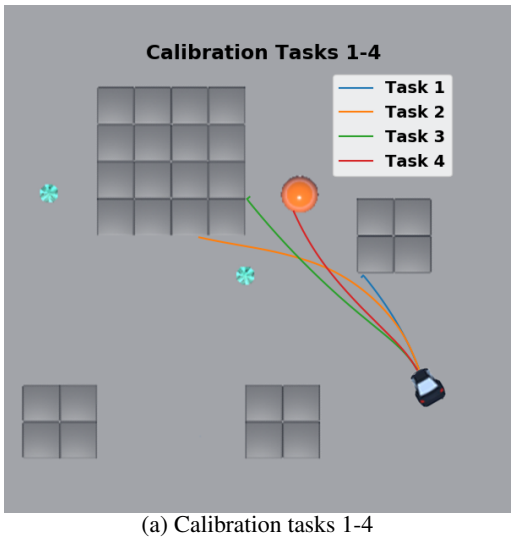
Table 4 lists the mean and standard deviations of the change in the embedding distance and the subjective metrics between the Feedback and No Feedback conditions. Study 3 employed the same subjective metrics as Study 2: trust, team fluency, understanding, workload, likeability, and perceived intelligence. To compare between conditions, we utilized either a one-tailed t-test, if the model passed normality and homoscedasticity assumptions or a one-tailed Wilcoxon Signed Rank test, a non-parametric test. We employed one-tailed tests because we hypothesized that the Feedback condition would be better on all metrics (higher for change in embeddings, likeability, perceived intelligence, trust, fluency, and understanding and lower for workload) than No Feedback.

As stated in the main paper, the amount that a person's embedding improved in the over-/under-correcting dimension, $\Delta\epsilon_{o/u}$, was significantly higher in the Feedback condition compared to No Feedback ($p = .006$). Although not significant, the amount that a person's embedding changed in the anticipatory/delayed, $\Delta\epsilon_{a/d}$, dimension was an improvement in the Feedback condition and got worse in the No Feedback condition. Additionally, the sum of these dimensions, $\Delta\epsilon_{o/u+a/d}$, was significantly improved in the Feedback condition compared to No Feedback ($p = .009$).

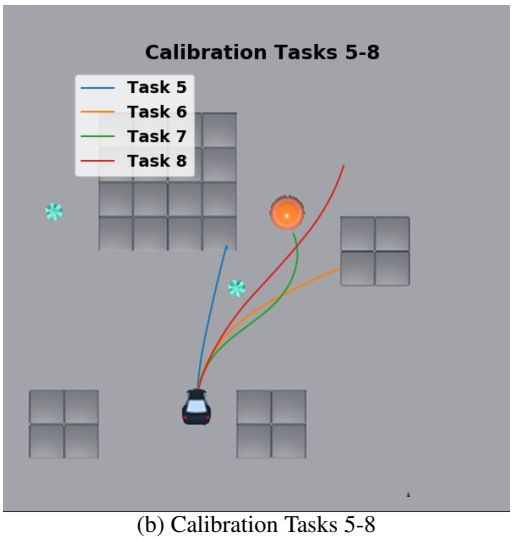
In terms of subjective metrics, we find the Feedback condition to be significantly lower in terms of workload compared the No Feedback condition ($p = .039$). Also, we find that participant's trust increased significantly more in the Feedback condition compared to No Feedback ($p = .019$). We additionally find that participants' perceived intelligence of the robot is trending towards being significantly higher for the Feedback condition compared to No Feedback ($p = .081$). Lastly, while not significant, participants' perceived team fluency and understanding increased more in the Feedback condition versus the No Feedback condition.

Table 4: This table shows the mean, (standard deviation), and test statistics of the subjective metrics, $\Delta\epsilon_{o/u+a/d}$, $\Delta\epsilon_{o/u}$, and $\Delta\epsilon_{a/d}$ for Study 3. Δ Trust, Δ Fluency, and Δ Understanding describe the change in Trust, Fluency, and Understanding respectively between the first and last round.

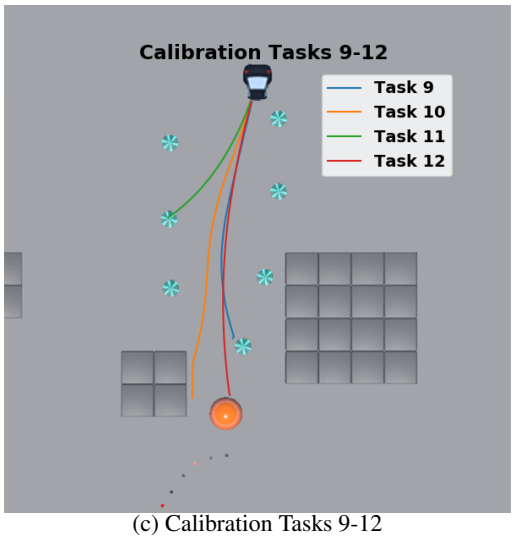
	Feedback	**No Feedback**	Test Statistic	p-value
$\boldsymbol{\Delta\epsilon_{o/u+a/d}}$	0.17 (0.35)	-0.05 (0.37)	$t(57.8) = 2.45$	$\boldsymbol{p = .009}$
$\boldsymbol{\Delta\epsilon_{o/u}}$	0.16 (0.28)	0.004 (0.19)	$t(52.0) = 2.62$	$\boldsymbol{p = .006}$
$\boldsymbol{\Delta\epsilon_{a/d}}$	0.01 (0.36)	-0.05 (0.33)	$t(57.3) = .724$	$p = .236$
Workload	44.3 (15.5)	51.4 (15.3)	$t(58.0) = -1.79$	$\boldsymbol{p = .039}$
Likeability	7.14 (1.67)	7.24 (1.65)	$t(58.0) = -.233$	$p = .592$
Intelligence	7.09 (1.14)	6.62 (1.40)	$t(55.7) = 1.42$	$p = .081$
Δ **Trust**	0.79 (0.75)	0.39 (0.51)	$Z = -2.07$	$\boldsymbol{p = .019}$
Δ **Fluency**	0.49 (0.56)	0.31 (0.54)	$t(57.9) = 1.25$	$p = .108$
Δ **Understanding**	0.49 (0.53)	0.42 (0.86)	$Z = -.790$	$p = .430$



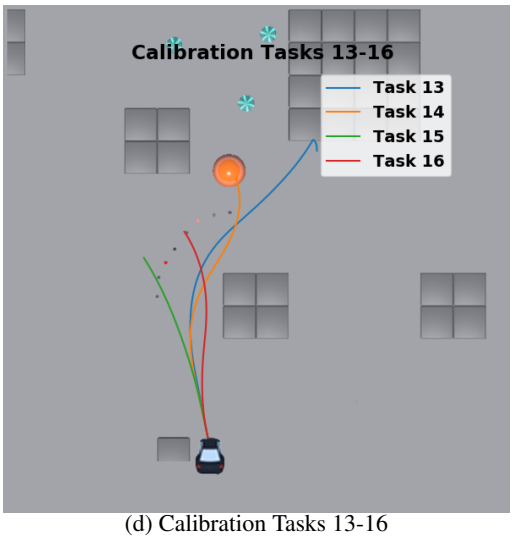
(a) Calibration tasks 1-4



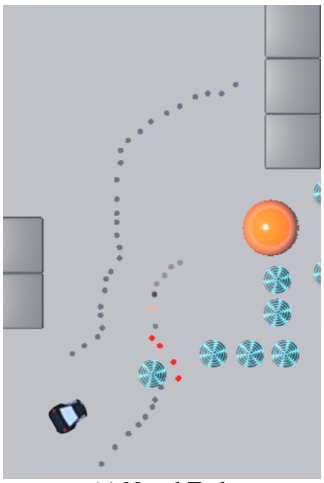
(b) Calibration Tasks 5-8



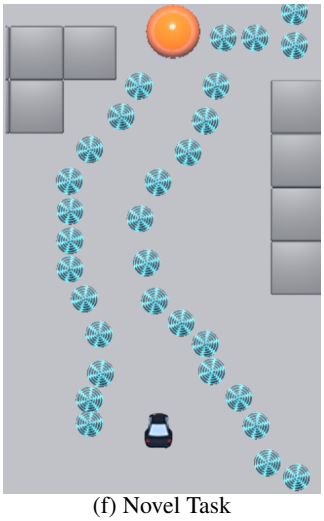
(c) Calibration Tasks 9-12



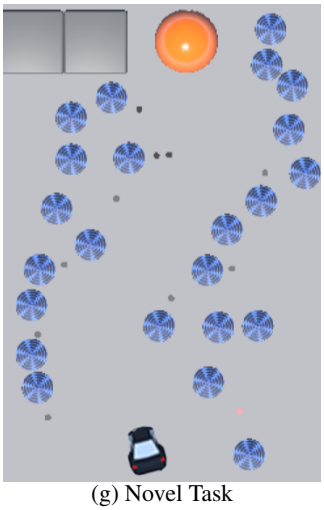
(d) Calibration Tasks 13-16



(e) Novel Task



(f) Novel Task



(g) Novel Task

Figure 6: This figure depicts the calibration tasks and novel tasks in the study. Figures 6a-6d show the calibration tasks. The car is the starting location and the orange ball is the goal location. The rest of the objects in the scene are obstacles. Each line represents one of the pre-recorded rollouts, which are a mix of successful and unsuccessful trajectories. Figures 6e-6g show the environment for the novel tasks. There are no rollout lines because the trajectories were dependent on participant input. The calibration tasks are simpler, have less obstacles, and less turns than the novel tasks.

9 Model Assumptions (Table 5)

Table 5: This table lists the statistical models and tests utilized in our analysis. The dependent variable (DV) and independent variable (IV) are specified for each model. We tested for normality using the Shapiro-Wilk test. When the IV is categorical, we employed Levene's test for homoscedasticity, otherwise, we employed the Breusch-Pagan test. If the model did not pass normality or homoscedasticity, then we used a non-parametric version of the statistical test.

Study 1				
DV	IV	Test	Normality	Homoscedasticity
$\epsilon_{o/u}^{(i)}$	Cooperative, $i = 1, 4$	Friedman's	N/A	N/A
$\epsilon_{o/u}^{(i)}$	Adversarial, $i = 1, 4$	rANOVA	$p = .424$	$p = .149$
$\epsilon_{o/u}^{(i)}$	None, $i = 1, 4$	rANOVA	$p = .706$	$p = .856$
DTW Round i	Cooperative, $i = 1, 4$	Friedman's	N/A	N/A
DTW Round i	Adversarial, $i = 1, 4$	Friedman's	N/A	N/A
DTW Round i	None, $i = 1, 4$	Friedman's	N/A	N/A
$\Delta\epsilon_{o/u}$	Condition	ANOVA	$p = .547$	$p = .931$
DTW Last Round	Condition	ANOVA	$p = .179$	$p = .855$
Workload	Condition	ANOVA	$p = .598$	$p = .454$
Likeability	Condition	ANOVA	$p = .770$	$p = .459$
Intelligence	Condition	ANOVA	$p = .571$	$p = .632$
ΔTrust	Condition	ANOVA	$p = .907$	$p = .925$
ΔTeam Fluency	Condition	ANOVA	$p = .457$	$p = .558$
Study 2				
DV	IV	Test	Normality	Homoscedasticity
$\epsilon_{o/u+a/d}^{(i)}$	Simultaneous, $i = 1, 5$	rANOVA	$p = .092$	$p = .826$
$\epsilon_{o/u+a/d}^{(i)}$	Greedy, $i = 1, 5$	rANOVA	$p = .167$	$p = .723$
$\epsilon_{o/u+a/d}^{(i)}$	None, $i = 1, 5$	rANOVA	$p = .081$	$p = .194$
$\Delta\epsilon_{o/u+a/d}$	Condition	ANOVA	$p = .708$	$p = .614$
$\Delta\epsilon_{o/u}$	Condition	ANOVA	$p = .100$	$p = .448$
$\Delta\epsilon_{a/d}$	Condition	ANOVA	$p = .565$	$p = .589$
Workload	Condition	ANOVA	$p = .573$	$p = .373$
Likeability	Condition	ANOVA	$p = .752$	$p = .587$
Intelligence	Condition	ANOVA	$p = .238$	$p = .453$
ΔTrust	Condition	ANOVA	$p = .290$	$p = .368$
ΔTeam Fluency	Condition	ANOVA	$p = .091$	$p = .201$
ΔUnderstanding	Condition	ANOVA	$p = .782$	$p = .883$
Study 3				
DV	IV	Test	Normality	Homoscedasticity
Final Distance	Condition	Wilcoxon*	N/A	N/A
Average Distance	$\epsilon_{o/u+a/d}$	Spearman's Correlation	N/A	N/A
$\Delta\epsilon_{o/u+a/d}$	Condition	t-test*	$p = .578$	$p = .848$
$\Delta\epsilon_{o/u}$	Condition	t-test*	$p = .402$	$p = .058$
$\Delta\epsilon_{a/d}$	Condition	t-test*	$p = .193$	$p = .540$
Workload	Condition	t-test*	$p = .814$	$p = .951$
Likeability	Condition	t-test*	$p = .057$	$p = .557$
Intelligence	Condition	t-test*	$p = .697$	$p = .416$
Δ Trust	Condition	Wilcoxon*	N/A	N/A
Δ Team Fluency	Condition	t-test*	$p = .732$	$p = .293$
Δ Understanding	Condition	Wilcoxon*	N/A	N/A

*Test is one-tailed.

References

[1] S. Ross, G. J. Gordon, and J. A. Bagnell. No-Regret Reductions for Imitation Learning and Structured Prediction. In *14th Internation Conference on Artificial Intelligence and Statistics*, Fort Lauderdale, FL, 2011.

[2] M. Laskey, C. Chuck, J. Lee, J. Mahler, S. Krishnan, K. Jamieson, A. Dragan, and K. Goldberg. Comparing human-centric and robot-centric sampling for robot deep learning from demonstrations. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 358–365, 2017. doi:10.1109/ICRA.2017.7989046.

[3] J. Spencer, S. Choudhury, M. Barnes, M. Schmittle, M. Chiang, P. Ramadge, and S. Srinivasa. Learning from Interventions: Human-robot interaction as both explicit and implicit feedback. *Robotics: Science and Systems XVI*, 2020. doi:10.15607/rss.2020.xvi.055.

[4] M. Laskey, S. Staszak, W. Y. S. Hsieh, J. Mahler, F. T. Pokorny, A. D. Dragan, and K. Goldberg. SHIV: Reducing supervisor burden in DAgger using support vectors for efficient learning from demonstrations in high dimensional state spaces. *Proceedings - IEEE International Conference on Robotics and Automation*, 2016-June:462–469, 2016. ISSN 10504729. doi:10.1109/ICRA.2016.7487167.

[5] M. Kelly, C. Sidrane, K. Driggs-Campbell, and M. J. Kochenderfer. HG-DAgger: Interactive imitation learning with human experts. *2019 International Conference on Robotics and Automation (ICRA)*, pages 8077–8083, 2019.

[6] D. H. Grollman, Aude, and G. Billard. Robot learning from failed demonstrations. *International Journal of Social Robotics*, 4(4):331–342, 2012.

[7] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza. Power to the people: The role of humans in interactive machine learning. *AI Magazine*, 35(4):105–120, 2014. ISSN 07384602. doi:10.1609/aimag.v35i4.2513.

[8] L. Chen, R. R. Paleja, and M. C. Gombolay. Learning from suboptimal demonstration via self-supervised reward regression. In *Conference on Robot Learning*, 2020.

[9] D. Brown, W. Goo, P. Nagarajan, and S. Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 783–792. PMLR, 2019.

[10] D. S. Brown, W. Goo, and S. Niekum. Better-than-demonstrator imitation learning via automatically-ranked demonstrations. In L. P. Kaelbling, D. Kragic, and K. Sugiura, editors, *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pages 330–359. PMLR, 30 Oct–01 Nov 2020. URL `https://proceedings.mlr.press/v100/brown20a.html`.

[11] B. Burchfiel, C. Tomasi, and R. Parr. Distance minimization for reward learning from scored trajectories. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30, 2016. doi:10.1609/aaai.v30i1.10411.

[12] V. Myers, E. Biyik, N. Anari, and D. Sadigh. Learning multimodal rewards from rankings. In A. Faust, D. Hsu, and G. Neumann, editors, *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 342–352. PMLR, 2022.

[13] M. Valko, M. Ghavamzadeh, and A. Lazaric. Semi-supervised apprenticeship learning. In M. P. Deisenroth, C. Szepesvári, and J. Peters, editors, *Proceedings of the Tenth European Workshop on Reinforcement Learning*, volume 24 of *Proceedings of Machine Learning Research*, pages 131–142, Edinburgh, Scotland, 2013. PMLR.

[14] M. Kaiser, H. Friedrich, and R. Dillmann. Obtaining good performance from a bad teacher. In *In International Conference on Machine Learning, Workshop on Programming by Demonstration*, 1995.

[15] K. Menda, K. Driggs-Campbell, and M. J. Kochenderfer. Ensembledagger: A bayesian approach to safe imitation learning. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5041–5048, 2019. doi:10.1109/IROS40897.2019.8968287.

[16] M. L. Schrum, E. Hedlund, and M. C. Gombolay. Improving Robot-Centric Learning from Demonstration via Personalized Embeddings. *arXiv*, 2021. 2110.03134.

[17] M. Cakmak and L. Takayama. Teaching people how to teach robots: The effect of instructional materials and dialog design. In *ACM/IEEE International Conference on Human-Robot Interaction*, pages 431–438. IEEE Computer Society, 2014. ISBN 9781450326582. doi:10.1145/2559636.2559675.

[18] A. Sena and M. Howard. Quantifying teaching behavior in robot learning from demonstration. *International Journal of Robotics Research*, 39:54–72, 1 2020. ISSN 17413176. doi:10.1177/0278364919884623.

[19] M. L. Schrum, E. Hedlund-Botti, N. Moorman, and M. C. Gombolay. MIND MELD: Personalized Meta-Learning for Robot-Centric Imitation Learning. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '22, pages 157–165. IEEE Press, 2022.

[20] R. Paleja and M. Gombolay. Inferring personalized bayesian embeddings for learning from heterogeneous demonstration. *arXiv*, 2019. ISSN 23318422.

[21] S. Salvador and P. Chan. Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal.*, 11(5):561–580, 2007. ISSN 1088-467X.

[22] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 2180–2188, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

[23] J.-Y. Jian, A. Bisantz, and C. Drury. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics*, 4:53–71, 2000. doi:10.1207/S15327566IJCE0401_04.

[24] G. Hoffman. Evaluating Fluency in Human-Robot Collaboration. *IEEE Transactions on Human-Machine Systems*, 49(3):209–218, 2019. ISSN 21682291. doi:10.1109/THMS.2019.2904558.

[25] S. G. Hart and L. E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In P. A. Hancock and N. Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland, 1988. doi:https://doi.org/10.1016/S0166-4115(08)62386-9.

[26] C. Bartneck, E. Croft, and D. Kulic. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1(1):71–81, 2009. doi:10.1007/s12369-008-0001-3.

[27] N. L. Robinson, T.-N. Hicks, G. Suddrey, and D. J. Kavanagh. The robot self-efficacy scale: Robot self-efficacy, likability and willingness to interact increases after a robot-delivered tutorial. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 272–277, 2020. doi:10.1109/RO-MAN47096.2020.9223535.