Learning from Unreliable Human Action Advice in Interactive Reinforcement Learning

Lisa Scherf[1,2], Cigdem Turan[1,2], Dorothea Koert[1,2]

***Abstract*— Interactive Reinforcement Learning (IRL) uses human input to improve learning speed and enable learning in more complex environments. Human action advice is here one of the input channels preferred by human users. However, many existing IRL approaches do not explicitly consider the possibility of inaccurate human action advice. Moreover, most approaches that account for inaccurate advice compute trust in human action advice independent of a state. This can lead to problems in practical cases, where human input might be inaccurate only in some states while it is still useful in others. To this end, we propose a novel algorithm that can handle state-dependent unreliable human action advice in IRL. Here, we combine three potential indicator signals for unreliable advice, i.e. consistency of advice, retrospective optimality of advice, and behavioral cues that hint at human uncertainty. We evaluate our method in a simulated gridworld and in robotic sorting tasks with 28 subjects. We show that our method outperforms a state-independent baseline and analyze occurrences of behavioral cues related to unreliable advice.**

I. Introduction

Interactive Reinforcement Learning (IRL) [1], [2] provides a potentially powerful approach to enable self-improvement of future humanoid robots through direct interactions with their environment and human teachers. In contrast to classic Reinforcement Learning (RL) [3], in IRL human feedback and advice can help to increase learning speed, improve sampling efficiency, and enable learning in more complex environments by exploiting human prior knowledge [4]. However, one limitation of many existing IRL approaches is the assumption that human input is always useful and correct [4]. In reality, human teachers might not always be able to provide such idealized input and human feedback or action advice might be partially incorrect for specific states, e.g. when the human teacher has a limited understanding of parts of a task or insufficient knowledge of an underlying state [5], [6]. A crucial step toward real-world applicability of IRL is therefore the development of algorithms that are capable to detect and handle potentially unreliable human input.

In this paper, we introduce a new IRL algorithm for learning from unreliable action advice (LUNAA), where we combine three different indicators for potentially inaccurate human input. Specifically, LUNAA computes the trustworthiness of action advice based on the consistency of the history of advice in a particular state, the retrospective optimality of human advice given the overall received environmental rewards, and implicit behavioral cues for human uncertainty, e.g. human response times and facial dynamics. In contrast to the majority of related approaches [6]–[8], we compute the trust in human action advice state-dependently. This allows LUNAA to only discard inaccurate action advice for particular states, while still profiting from correct human input in other states. The use of state-dependent trust has also been proposed for human input in the form of evaluative feedback before in [9], however, in our approach we apply it to human action advice, which was shown to be one of the preferred human input sources in IRL [6], [10].

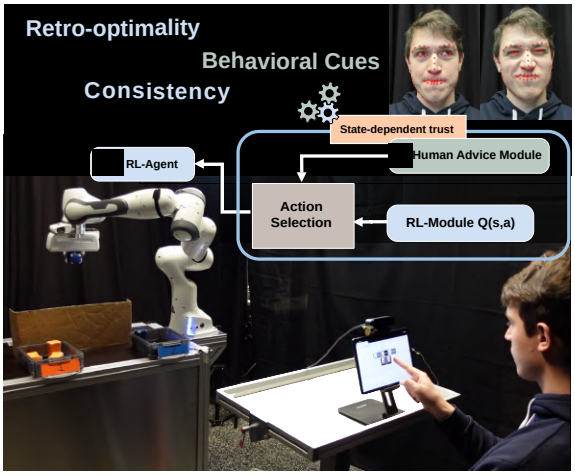

Fig. 1: We propose a new IRL algorithm that combines three indicators, i.e. consistency, retrospective optimality, and behavioral cues, into a measure for state-dependent trust to handle potentially unreliable action advice.

Experiments in a grid world and a robotic sorting task show advantages of our approach over a state-independent decrease of trust in advice, as it was proposed e.g. in [8] and [6]. Additionally, we evaluate response time and facial dynamics as behavioral cues for insecure and potentially inaccurate advice in robotic sorting tasks with 28 participants.

II. Related work

There are several ways of incorporating human input into RL [4]. One main difference in existing approaches is the form in which humans can provide input, e.g. as evaluative feedback after action execution [11], [12] or proactive guidance signals in form of action advice [13]. However, the majority of IRL approaches assume the human input to be useful and correct [4], which may not always be the case in real applications [5], [6], [14]. Some approaches account for false sensor detection of human input but assume the underlying human input to be optimal [15]–[17]. Only a few works try to explicitly model strategies behind human input [14], [18], [19] or consider cases where the human input itself might be inaccurate or suboptimal [6], [8], [20]–[23].

Griffith et al. [7] proposed the ADVICE algorithm where they follow a Bayesian approach to estimate a consistency parameter for human feedback. However, their consistency

This Work was funded by German Federal Ministry of Education and Research (project 01IS20045). [1] Interactive AI & Cognitive Models for AI interaction, [2] Centre for Cognitive Science, TU Darmstadt, Germany
lisa_katharina.scherf@tu-darmstadt.de

parameter is estimated over the whole state space and can not account for state-dependent errors in human feedback signals. Other works [6], [8], [24] introduced heuristic time-dependently decreasing trust in human feedback, following the assumption that over time the policy starts to generate better output. However, these approaches also make no difference in trust for human input over different states.

The use of ensembles of experts as feedback sources to compute the reliability of human inputs has also been proposed [21], [23]. However, these approaches are limited by the availability of multiple human teachers and also do not consider state-dependent inaccuracies of specific teachers. Other approaches compute trust in a learned policy depending on the current DQN loss function and compare advice and learned Q-values to account for inconsistencies, combine a policy learned from human feedback with the agent's policy based on the match of human feedback with multiple stored policies from the environmental reward function [22], or use an Expectation-Maximization approach for learning from uncertain input [25]. Learning from inattentive teachers [5] and explicit considerations for state-dependent inaccuracies have been proposed by [9] and [5]. In particular, [5] discusses that the assumptions of underlying patterns to incorrect human input might depend on misunderstandings of tasks or robot capabilities and that it is advantageous for an agent to learn to distinguish such patterns instead of trusting all human feedback equally. In the REPaIR algorithm [9], they propose an approach that can learn whether to keep, invert, or discard human feedback given in binary form, dependent on the overall achieved environmental reward for specific state-action pairs. Our indicator retro-optimality for human action advice follows this basic idea of the REPaIR approach. However, while [9] is designed for evaluative feedback signals only, our approach focuses on handling state-dependent inaccuracies in human action advice signals. We focus here on action advice since recent studies have shown that humans prefer to not only give evaluative feedback but also want to give active guidance [6], [10]. In contrast to the existing approaches that used comparison with environmental rewards [9], quality of the current policy [20], or consistency of human input [7], [20], [25] as indicators of trust, we propose a third indicating source to help an agent to detect potentially inaccurate human advice. Our use of behavioral models as an additional indicator is inspired by findings that humans also learn over time how to perceive and detect signals for uncertainty in other humans [26]. Human uncertainty occurs in different forms [27] and it was shown before in human-computer interaction literature that it is possible to detect uncertain input, confusion, or lies of a human user from behavioral and physiological signals [28]–[31]. In this paper, we explore facial dynamics and response times as behavioral cues. Facial dynamics have been explored as rewards in IRL [32] but not as an indication of potentially wrong input. We also found no work that explicitly considers response times for state-dependent trust in advice for IRL.

III. Learning from Unreliable Action Advice

In this section, we introduce LUNAA, a new IRL algorithm that learns from unreliable human action advice. We model the problem as a Markov Decision Process (MDP), where for action a in a state s, the agent transitions to the next state s' and receives a reward r. The goal is to learn an optimal policy $\pi^*(s)$ that maximizes the expected cumulative reward. For the experiments in this paper, we use tabular Q-learning and update the Q-function according to

$$Q(s,a) \leftarrow Q(s,a)+\alpha(r+\gamma \max_{a'} Q(s',a')-Q(s,a)), \quad (1)$$

where α is the learning rate and γ the discount factor. Here, we used $\gamma = 0.98$ and a fixed learning rate $\alpha = 0.1$.

In addition, we learn a model of human action advice $H(s,a)$ to predict which action the human would most likely suggest for a given state s. Such a learned model of human action advice can be beneficial since human input might be sparse and humans tend to give less advice over time [33]. A human advice module is comparable to learning a model for human reward based on human feedback as proposed in [2] and has also been suggested in [6]. However, under the assumption that human advice a_h can be partially incorrect, $H(s,a)$ can deviate from the optimal policy $\pi^*(s)$ for the given reward function R. Therefore, we additionally define a function $t_h(s,a_h)$ that assigns a trust in $H(s,a)$ in the range $[0,1]$ for each state and action. In contrast to prior work [6]–[8], we propose a state-dependent formulation for trust. Similar to Kessler Faulkner et al. [9], we assume that the correctness of human action advice often depends on the current state rather than simply improving or worsening over time. In particular, a human might only be inaccurate in some states due to partly misunderstanding a task, while still being able to provide useful advice in other states.

In the following, we explain three indicators that we use to compute the state-dependent trust, i.e. consistency $c(s,a_h)$ in Section III-A, retrospective optimality $o(s,a_h)$ in Section III-B, and behavioral cues $b(s,a_h)$ in Section III-C. In Section III-D we show how we learn the human advice module and in Section III-E we explain how to use the state-dependent trust to combine the learned advice module and the Q-function. Algorithm 1 summarizes the approach.

A. Consistency

One indicator to detect unreliable human input is the consistency of input over time [34]. To detect states with inconsistent action advice, we store the last N_Φ advised actions a_h for each state s in a state-specific advice history Φ_s. Based on this advice history we then compute a consistency factor $c(s,a_h)$ for each state and action pair

$$c(s,a_h) = \max(0, ((N_{a_h}/N_\Phi) - 0.5)/0.5), \quad (2)$$

where N_{a_h} is the number of occurrences of the advised action a_h in the action history $\Phi(s)$. For the experiments in this paper, we limit the maximum number of stored actions in the history of a state to 10. Whenever this limit is exceeded for a state we discard the oldest stored action from $\Phi(s)$.

B. Retrospective Optimality

As a second indicator for potentially unreliable advice we propose retrospective optimality (retro-optimality), that is if human advice leads to one of the lower seen cumulative rewards, it is more likely suboptimal. Similarly, Kessler

Algorithm 1 LUNAA

Require: max number of episode steps M, max episodes E
init Q-table $Q[s,a]$=0 $\quad \forall s, a$ if a possible in s, else $-\infty$
init visits per state $v[s] = 0 \quad \forall s$, episode counter $e = 0$
init $\gamma = 0.98$, $\varepsilon = 0.1$, $\alpha = 0.1$
while $e <$ max number episodes E **do**
  $s =$ random init state, episode steps counter $i = 0$
  while episode not *finished* **and** $i <$ M **do**
    ξ_e = [] empty list, r_ξ = 0, $v[s] = v[s] + 1$
    if action advice $\{a_h, b_i(s, a_h)\}$ given **then**
      store s, a_h into ξ_e
      Update $H(s,a)$, $b(s,a_h)$, $c(s,a_h)$, Eq.(9), (8), (2)
      decide if to reject a_h based on Eq.(14)
    else
      $a_h = \arg\max_a[H(s,a)]$
      Compute $t_h(s,a_h)$, t_q, $\beta(s)$, Eq.(10), (11), (12)
      p sample from uniform distribution
      if $p < (1 - \beta(s))$ **then**
        $a = a_h$
      else
        $a =$ choose ε-greedy action a from $Q[s,a]$
      end if
    end if
    execute action a, get reward r and next state s'
    $Q[s,a] = Q[s,a] + \alpha(r + \gamma \max_a Q[s',a] - Q[s,a])$
    $s = s'$, $r_\xi + r$, $i = i + 1$,
  end while
  $R_{\max}[(s,a_h)] = r_\xi$
  for all $(s_\xi, a_\xi) \in \xi_e$ **do**
    Update $o(s_\xi, a_\xi)$, Eq.(3)
  end for
  $e = e + 1$
end while

Faulkner et al. [9] use this intuition to estimate corrections to partially incorrect feedback over time. Here, we adapt their approach and computation of trust for action advice instead of feedback. During each episode, we save the state trajectory ξ together with the provided action advice a_h. At the end of each episode, the cumulative reward R_ξ is stored. For (s, a_h), if a higher R_ξ has not been seen, we save it as the highest reward for this state-action pair: $R_{\max}[(s,a_h)] = r_\xi$. Each episode the retro-optimality $o(s,a_h)$ is updated as

$$o(s,a_h) = \frac{R_{\max}[(s,a_h)] - \min(R_{\max})}{\max(R_{\max}) - \min(R_{\max})} \tag{3}$$

Retro-optimality can here be beneficial as soon as the agent has experienced ways to achieve a higher reward than when just following the current human action advice.

C. Behavioral Cues

As the third indicator for unreliable action advice, we propose the use of behavioral cues. This is inspired by the fact that also humans learn to detect uncertainty in other humans based on behavioral signals [26]. In comparison to consistency and retro-optimality, behavioral cues can potentially help to identify incorrect advice from the very first episode. Behavioral cues may help when incorrect advice correlates with human uncertainty. However, if there is a mismatch between the level of confidence and the correctness of the advice, behavioral cues might lead to a rejection of correct advice or acceptance of wrong advice (Fig. 2). In this case, the combination with additional indicators, i.e. inconsistency and retro-optimality, can be particularly beneficial. Different physiological and behavioral signals have been shown to relate to behavioral uncertainty in humans [28]–[31]. We focus here on response times and facial dynamics.

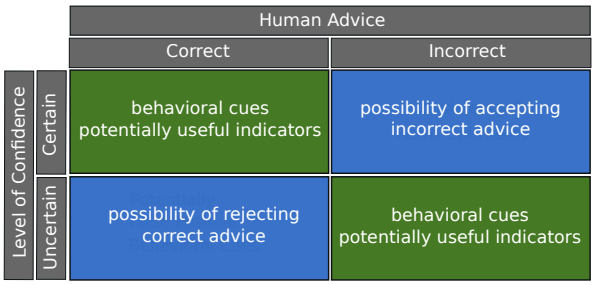

Fig. 2: Human advice can be correct or incorrect and given with a high or low level of confidence. Behavioral cues may help when incorrect advice correlates with uncertainty.

1) Response Time: Evidence suggests that a higher response time can indicate a higher uncertainty level of a user [28]. The response time $RT[ms]$ is here defined as the time between the presentation of all options of action advice to the user and the selection of one action a_h. Based on RT, we compute a certainty indicator $rt(s,a_h)$ between 1 (certain) and 0 (uncertain) for a state s and action advice a_h as

$$rt(s,a_h) = 1 - \min(1, \max(0, \frac{RT(s,a_h) - RT_{\min}}{RT_{\max} - RT_{\min}})) \tag{4}$$

where $RT_{\min}$ and $RT_{\max}$ are upper and lower thresholds on the response times.

2) Facial Behaviors: To estimate a human's level of uncertainty for given advice from facial behaviors, we use OpenFace [35] to extract Action Units (AUs), which encode regional facial movements and their intensities [36]. We consider the time window between enabling the advice GUI and input of the user at time t. We compute the average sum of absolute differences between the n frames for the subset of AUs recognized by OpenFace $FB = \frac{1}{n}\sum_{AUs}\sum_{i=t-n}^{t}|AU_i - AU_{i+1}|$. We hereby excluded AUs mostly related to blinking (AU05, AU07, AU45) after pilot data analysis. We compute a facial certainty indicator

$$f(s,a_h) = 1 - \min(1, \max(0, \frac{FB(s,a_h) - FB_{min}}{FB_{\max} - FB_{\min}})), \tag{5}$$

where $FB_{\min}$ and $FB_{\max}$ are thresholds between 0 and 1.

3) Combined behavioral indicator: We compare Linear Opinion Pool [37] with equal weights (LOP) and Independent Opinion Pool [38] (IOP) to combine behavioral cues

$$b_i^{\text{LOP}}(s,a_h) = 0.5 \cdot (rt_i(s,a_h) + f_i(s,a_h)) \tag{6}$$

$$b_i^{\text{IOP}}(s,a_h) \propto rt_i(s,a_h) \cdot f_i(s,a_h). \tag{7}$$

In Section IV-B.2 we discuss the effects of these different combinations. Since behavioral signals for an action in a state might vary over time we compute a weighted average

$$b(s,a_h) \leftarrow (1 - \alpha_b) b(s,a_h) + \alpha_b b_i(s,a_h), \tag{8}$$

where α_b can be seen as a learning rate and is set to $1/N_\Phi$ for the experiments in this paper where N_Φ is the number of seen action advice with a maximum of 10.

D. Human-Advice-Module

As also proposed in [6], we learn a human advice module $H(s, a_h)$ to account for the fact that a human might only provide sparse action advice or decrease the amount of advice in the same state over time. For the experiments in this paper, we use a simplified function approximator similar to a tabular Q-function. Here, we initialize all entries with zero, and whenever an action a_h is advised in a state s with behavioral indicator behavioral $b_i(s, a_h)$ we update $H(s, a_h)$

$$H(s, a_h) = H(s, a_h) + \min(0.1, b_i(s, a_h)), \tag{9}$$

that gives more weight to advice with higher certainty. For using LUNAA without behavioral cues we set $b_i(s, a_h) = 1$.

E. State-dependent Action Selection Module

In Sections III-A-III-C we proposed three indicators for unreliable human advice. We combine all three indicators into an estimate of trust into human advice $t_h(s, a_h)$

$$t_h(s, a_h) = \min(c(s, a_h), o(s, a_h), b(s, a_h)). \tag{10}$$

This is a conservative combination, where we always distrust human advice if one of the indicators considers it distrustful. In addition, we compute the trust in the agent's own policy, similar as in [6], [20]. This we refer to as *self-confidence* hereafter. While $t_h(s, a_h)$ is computed state-dependent the self-confidence is state-independent and increases over time

$$t_q^e = \min(1.0, \max(0.0, e - e_{t_q,\text{start}})m_{t_q}), \tag{11}$$

where t_q^e is the agent's self-confidence in episode e, and $e_{t_q,\text{start}}$ and m_{t_q} denote starting point and slope of increase.

We use a combination of the state-dependent trust in human advice and the self-confidence for two things. In cases when no human input is provided, we use t_h and t_q to combine the $H(s, a)$ and $Q(s, a)$ for the agent's action selection policy. Motivated by [24], we use a shared control approach. We first compute $\hat{a}_{h,\max} = \arg\max_a H(s, a)$ and if multiple actions maximize $H(s, a)$, we decide for the one with maximal $t_h(s, \hat{a}_{h,\max})$. Then we compute

$$\beta(s) = \max((1.0 - t_h(s, a_{h,\max})), t_q^e), \tag{12}$$

as a state-dependent combination parameter for the shared control approach. Here, considering t_q^e ensures that even if none of the indicators (Sections III-A-III-C) recognizes an unreliable advice after a higher number of episodes the policy still takes over learning due to increasing trust in the agent's own Q-function. We use $\beta(s)$ as a probability to switch between the action computed by the human advice module and an ε-greedy policy based on the agents Q-function, except for states where no human advice has been given so far ($H(s, a_j) == 0\ \forall a_j$). Then we only consider the policy based on the Q-function

$$\begin{aligned} &\text{if not } H(s, a_j) == 0\ \forall a_j, \\ &\qquad P(a_\pi = \arg\max_a [H(s, a)]) = \beta(s), \\ &\text{else:}\quad P(a_\pi = \arg\max_a [Q(s, a)]) = 1 - \varepsilon(s), \\ &\qquad P(a_\pi = \text{random action}) = \varepsilon(s), \end{aligned} \tag{13}$$

where $\varepsilon = 1/\sqrt{v(s)}$ is the exploration parameter of an epsilon greedy policy based on number of state visits $v(s)$.

Second, we use the computed trust parameters in the case when a human actively gives action advice to give the agent the option to question and reject human advice

$$p_\text{reject} = \max((1.0 - t_h(s, a_h), t_q), \tag{14}$$

where with the probability p_reject human advice a_h gets rejected and the agent instead follows its learned policy.

IV. Experimental Evaluation

We present experiments in a grid world scenario (Sec. IV-A) and in robotic sorting tasks with 28 human participants (Sec. IV-B). We compare LUNAA against a state-independent linear increase of the agent's self-confidence (*T-SC*) as it was proposed e.g. in [6], [24]. Moreover, in the robotic tasks, we analyze how task-related uncertainties influence response times and facial dynamics of participants.

A. Gridworld with Simulated Human Input

Fig. 3(a) shows the used grid world with 4x5 states and four actions: up, down, left, and right. An episode ends if the agent reaches the goal (r=100), goes into a fire (r=-100), or exceeds the maximum number of 15 steps (r=0). In all states, which do not end the episode, the agent receives a reward of r=-1. Advice is given throughout the first 20 episodes. To simulate partially incorrect advice, the agent receives correct, useful feedback at all states except the purple state. In the purple state, we either simulate inconsistent wrong advice that is a random choice among all sub-optimal actions ('up', 'down', 'left') or consistent wrong advice that is 'up', to analyze the interplay of the chosen three indicators for different types of advice. We compare our method with a state-independent linear increase over time for a self-confidence (*T-SC*) as proposed in [6], [24]. Fig. 3(b) shows that when using a human advice module useful human advice speeds up learning (cyan), whereas for partially incorrect inconsistent advice learning fails (orange). Therefore, being able to distinguish reliable and unreliable human advice could be beneficial in this scenario. Fig. 4(a) shows that *T-SC* can speed up learning since over time the learned human advice function or human input itself can be questioned. However, if this self-confidence increases too early useful advice is not fully exploited and a later increase may result in following partially incorrect advice for a longer time. To tackle these problems, in contrast to T-SC in our approach (LUNAA) we introduce a state-dependent trust in human advice. Fig. 3(c) visualizes the results with mean and standard deviation over 50 evaluation runs and 100 experiments. A Kruskal Wallis test ($\chi^2 = 193.7$, $p < 0.001$) and posthoc Conover's show that for inconsistent wrong advice for the combination of our indicators consistency and retro-optimality (*LUNAA-CR*), we converge significantly faster than *T-SC* ($p < 0.001$) and *policy only* without human advice ($p < 0.001$). For consistent wrong advice, *LUNAA-CR* shows no discernible advantage over *T-SC* in learning (Fig. 3(d)). In such cases, we consider behavioral cues as additional indicators for human uncertainty particularly useful. We simulate behavioral certainty with $b \sim [1.0, 0.8]$ for certain advice and $b \sim [0.0, 0.2]$ for uncertain/incorrect

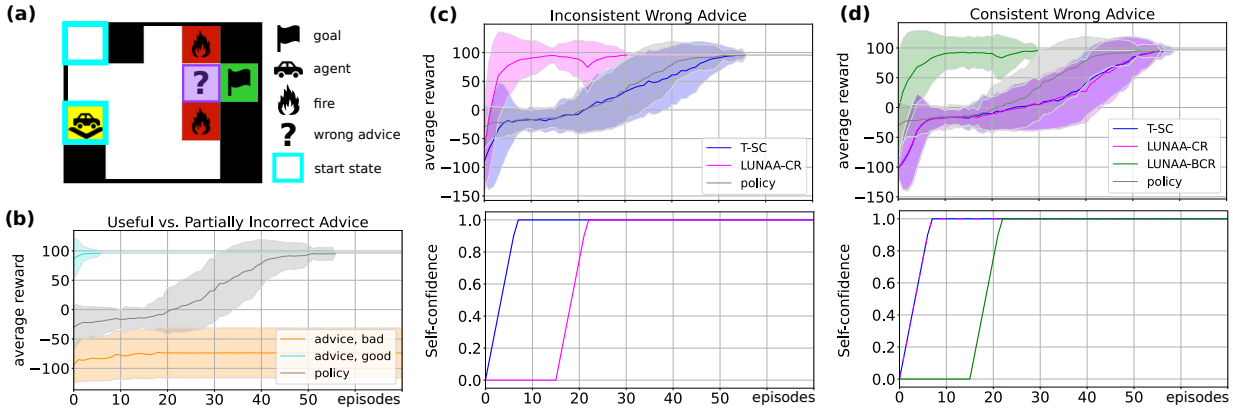

Fig. 3: (a) Gridworld scenario, (b) influence of optimal (cyan) versus partially incorrect advice (orange), (c) for inconsistent wrong advice *LUNAA-CR* learns significantly faster than *T-SC*, (d) benefit of *LUNAA-BCR* for consistent wrong advice. All plots show mean and standard deviation over 50 evaluation runs and 100 experiments

advice. The probability of false positives and negatives is set to 0.05. We compare our approach including behavioral models (*LUNAA-BCR*) against *LUNAA-CR* and *T-SC*. A Kruskal Wallis test ($\chi^2 = 239.3$, $p < 0.001$) and posthoc Conover's show that *LUNAA-BCR* converges significantly faster than *T-SC* ($p < 0.001$), *LUNAA-CR* ($p < 0.001$), and *policy only* ($p < 0.001$). *T-SC* and *LUNAA-CR* show no significant differences in convergence for consistent wrong advice ($p = 0.605$) and converge significantly slower than *policy only* ($p < 0.001$). In Fig. 4 we test the robustness of (a) *T-SC* and (b) *LUNAA-BCR* to variations in self-confidence parameters. The self-confidences were optimized separately for both methods before including variations. The plots reveal that *T-SC* is more sensitive to variations. *LUNAA-BCR* converges faster than *policy only* for all tested parametrizations. In Fig. 4(c) we evaluate the influence of inaccurate behavioral certainty estimations on learning with *LUNAA-BCR*. Up to a probability $P(\text{uncertain}|a_h \text{ is correct}) = 0.5$ (false negative), *LUNAA-BCR* converges faster than *policy only*. For a probability of $P(\text{certain}|a_h \text{ is incorrect}) = 0.5$ (false positive), *LUNAA-CR* is slower than learning without advice.

B. Robotic Sorting Tasks with Real Human Input

In the literature, we found a lack of experiments that investigate the occurrence of partially incorrect advice with real human subjects in IRL [6]. However, such experiments are crucial to better understand how to detect and handle such advice in real robotic tasks. We therefore conducted robotic experiments with 28 subjects (18 male, 10 female, age 18-35), where we investigate state-dependency of correctness in human advice and the occurrence of behavioral cues which could indicate state-dependent human uncertainty. The participants mostly reported a low level of prior experience with robots, in particular, 16 persons never or only once had contact with robots before and only two subjects reported more than 20 prior encounters. The experiments were approved by Ethikkommission of TU Darmstadt on 07/21/2021. For each subject, the experiment is divided into three parts (Ex.1A, Ex.1B, Ex.2), which differ in the extent to which subjects can assess their uncertainty about the correctness of given advice. In all experiments, the robot should sort objects correctly into boxes. The sorting criterion is related to the objects' weights (HIGH or LOW) which can not be accessed by the human but only by the robot's sensors in the moment when it lifts the objects. The reward is 10 if an object is sorted correctly and -10 if it is sorted incorrectly. The actions are GO-TO-OBJECT, GRASP, DROP, GO-TO-BOX-X. The state is defined by the weight on the robotic arm (EMPTY, HIGH, or LOW) and its gripper position (AT-HOME, AT-OBJECT, AT-BOX-X), resulting in 12 states for Ex.1A&B and 15 states for Ex.2. Advice is given over a tablet with a GUI shown in Fig. 5(c) and experimental setups for Ex1.A&B and Ex2. are shown in Fig. 5 (a) and (b).

1) Comparison against T-SC for Ex.1A&B: In Ex.1A&B the participants were told to help a robot sort objects colored blue or orange (Fig. 5 (a)). However, object colors did not correspond to the sorting criterion, i.e. object weights. Since the boxes had colored labels matching object colors and the sorting criterion was not communicated, the participants were likely influenced to get a false prior. This results in partially wrong advice (sorting objects in the wrong box), while still being able to provide useful advice for general task structure, e.g. first go to object then grasp. In Ex.1A, the participants did not receive feedback if an object was sorted correctly and are therefore expected to stick to their false prior about the sorting criterion. In Ex.1B, subjects received feedback after each episode and might therefore start to question their assumed sorting criterion. Fig. 6(a) shows that as intended these experimental settings resulted in partially incorrect human advice. In particular, the correctness of advice is state-dependent, where in states where they had to decide about which box to sort in (sorting) subjects gave a higher proportion of incorrect advice compared to states where they just advised about task-structure actions, i.e. go-to-object, grasp (non-sorting). Fig. 5(e) shows the comparison between *LUNAA-CR* and *T-SC* in Ex.1A&B. The plot shows the mean and standard deviation of the average reward for 50 evaluation runs and 50 experiments with different random seeds where we use the recorded human advice from Ex.1A and Ex.1B in the first 8 episodes and then continue learning with shared control policy between $Q(s,a)$ and $H(s,a)$. The use of consistency and retro-optimality as indicators for state-dependent incorrect advice speeds up learning significantly compared to using a time-dependent self-confidence only

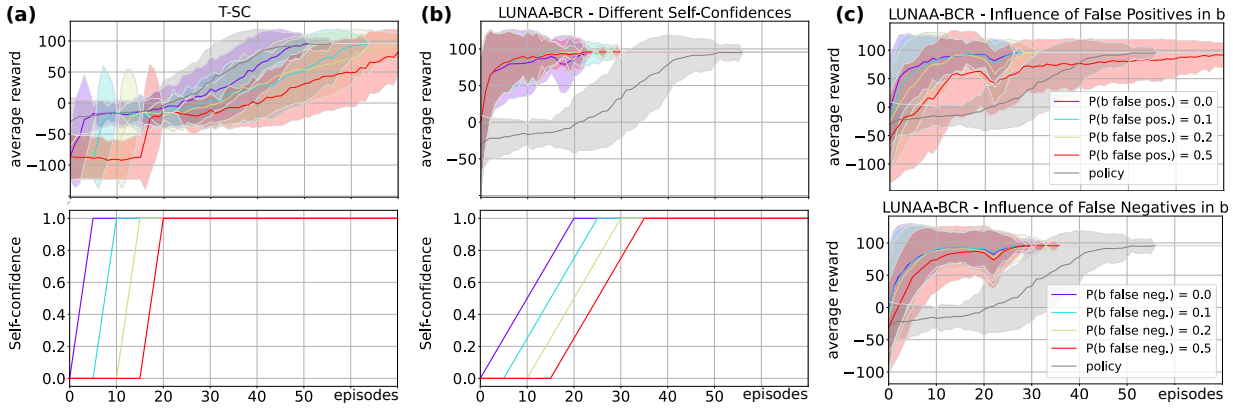

Fig. 4: Comparison of *T-SC* (a) and *LUNAA-BCR* (b), robustness to varying self-confidence parameters and evaluation of influence of errors in behavioral certainty estimates on *LUNAA-BCR* (c) (50 evaluation runs, 100 experiments)

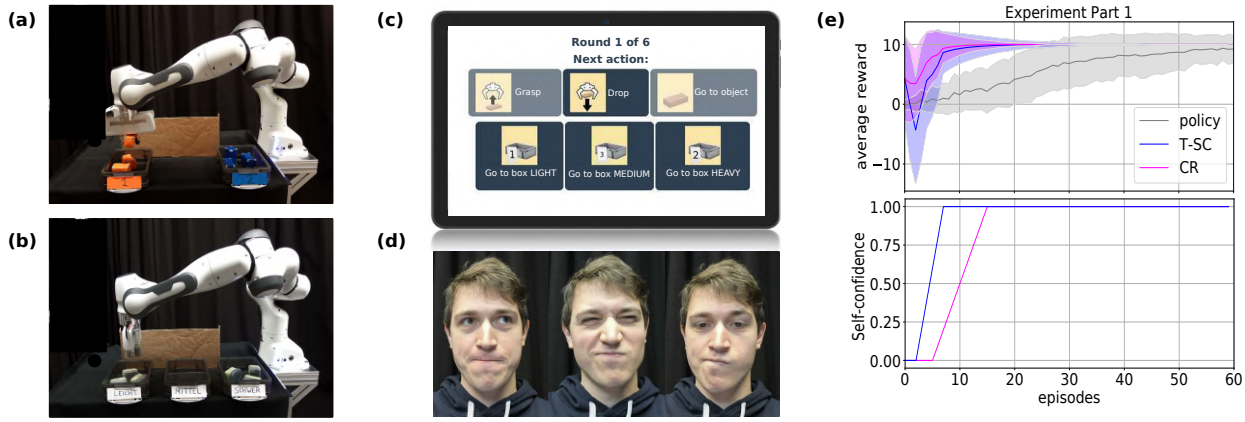

Fig. 5: Setup for Ex.1 (a) and Ex.2 (b), (c) shows GUI for action advice, (d) shows exemplary facial expressions. (e) shows for Ex.1 *LUNAA-CR* converged significantly faster than *T-SC* and *policy only* (50 evaluation runs, 50 experiments)

(Wilcoxon Signed Rank (WSR), $p < 0.001$, $Mdn = 4.6$ (*LUNAA-CR*) vs 5.3 (T-SC)) and using *policy only* (Mann-Whitney-U, $p < 0.001$).

2) Analysis for behavioral cues: In Ex. 1A&B LUNAA-CR, i.e. the indicators consistency and retro-optimality already showed advantages over the T-SC baseline (Sec. IV-B.1). As demonstrated in the gridworld we argue that using behavioral cues as an additional indicator could increase the benefit of LUNAA even further. To better understand occurrences of such behavioral cues and their relation with partially wrong human advice we provide here a pilot analysis for response times and facial dynamics as behavioral cues in Ex.1A&B and in an additional experiment, i.e. Ex.2.

Compared to Ex.1, in Ex.2 the underlying sorting criterion was more openly communicated to the subjects. They were briefed to assist the robot to sort objects in 3 boxes (labeled *light*, *medium*, *heavy*) according to weight (Fig. 5(b)). However, the exact thresholds for e.g. light or heavy were not communicated. The objects were filled with different material visible to the subjects, such that they could use their prior knowledge to get a sorting intuition. In a familiarization phase, the human advised two objects filled fully with feathers (*light*) and two filled fully with stones (*heavy*). This was followed by three experiment runs with six episodes, each. In each run, four objects were objects previously seen in the familiarization phase and two unknown objects filled with either a combination of stones and feathers or a varying number of screws, or a different amount of stones. Since those objects were not seen before, humans were expected to be less certain about where to sort them. In addition, the threshold was designed to potentially contradict their first intuitions about unknown objects resulting in state-dependent partially wrong advice (Fig 6(a)). In contrast to Ex.1, based on the state-dependent trust the robot could now also reject the advice and perform another action. At the end of each episode, the human received feedback on whether the task was successfully solved. Compared to Ex.1, in Ex.2 we expected subjects to be able to better assess their level of uncertainty for a particular state and object.

Over all robotic experiments, we evaluate if humans show uncertainty connected to unreliable advice and how this reflects in their response time (*RT*) and facial dynamics (*FB*) (Section III-C). In Fig. 6(d-f), we analyze *RT* and the resulting certainty indicator rt (Eq. 4) over 28 subjects for Ex.1A&B and Ex.2. In Fig. 6(d), we compare *RT* for advice directly related to sorting (purple) to non-sorting advice (orange: grasp, go-object). In Ex.1A the subjects did not receive feedback, so we expected them to be certain about sorting the objects according to color, despite being partially wrong. *RT* for sorting actions is here significantly lower (WSR, $z = 2.69$, $p < 0.01$) than for non-sorting, but with only a small difference in absolute values (Mdn 1.81s (non-

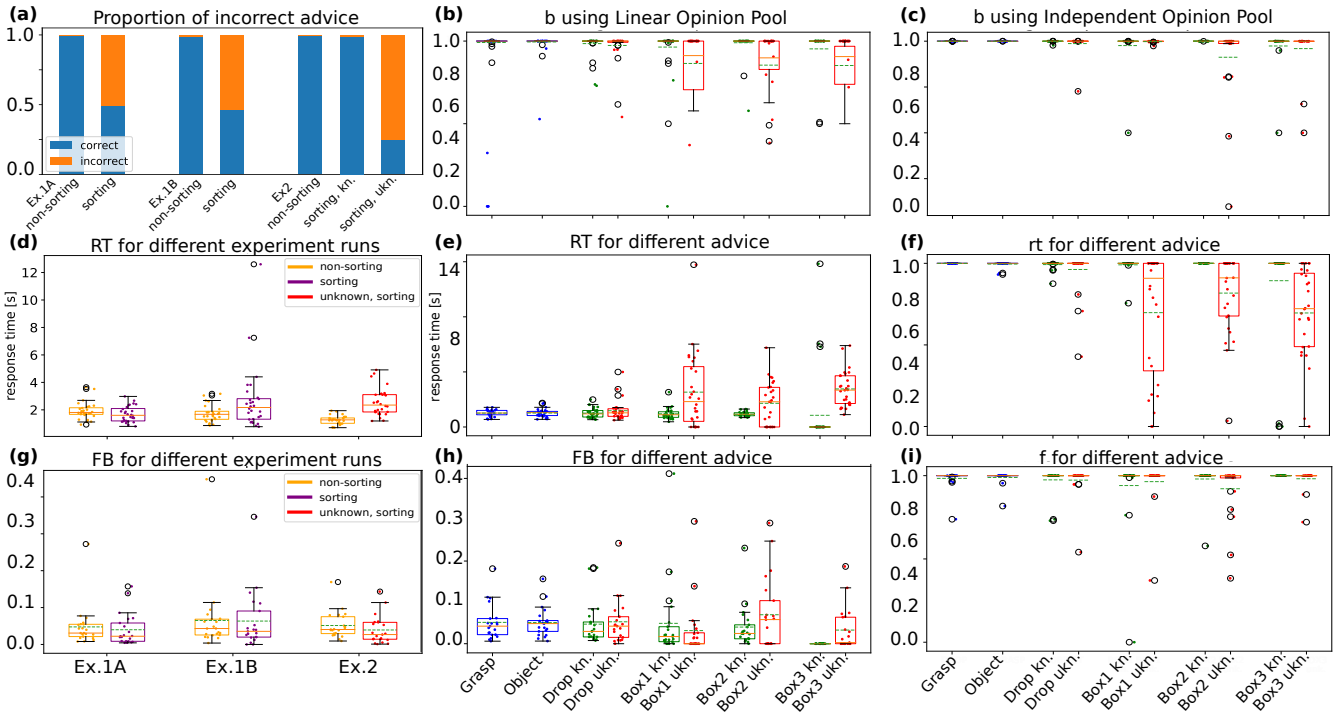

Fig. 6: We compare the occurrence of state-dependent incorrect advice (a), response times (d), and facial dynamics (g) for Ex.1A, Ex.1B, and Ex.2 for non-sorting advice (orange: grasp, go-to-object) vs. advice related to sorting (purple). We evaluate response times (d-f) and facial dynamics (g-i) and show median (orange), mean (dotted green line), lower/upper quartiles, and outliers. (e-f) and (h-i) compare non-sorting advice (blue) to choosing boxes for objects known from familiarization (green), or unknown objects (red). We compare a combination of rt and f into b-scores with LOP (b) and IOP (c).

sorting) vs 1.60s (sorting)). In Ex.1B participants received feedback at the end of each episode and might question their first assumed sorting criterion, potentially resulting in more behavioral uncertainty. *RT*s are significantly higher for sorting actions compared to non-sorting actions (WSR, $z = 2.23$, $p < 0.05$). In Ex.2 the sorting criterion was communicated to the participants such that they could use their prior knowledge about object weights. Here, *RT*s are significantly higher for sorting advice on unknown objects (red, $Mdn = 2.56$s) compared to non-sorting advice (orange, $Mdn = 1.27$s) (WSR, $z = 4.60$, $p < 0.001$). In Fig. 6(e) *RT* is shown for Ex.2 over all subjects for each action advice. We distinguish between non-sorting advice, sorting advice for known objects, and unknown objects. We expect subjects to be less certain when sorting unknown objects compared to objects known from the familiarization phase and non-sorting advice. *RT*s are significantly lower for sorting advice on known objects compared to non-sorting advice ($Mdn = 1.17$s sorting known vs. 1.27s non-sorting, WSR, $z = 2.71$, $p < 0.01$). In Fig. 6(f), we show rt (Eq. (4)) with $RT_{\min}$ and $RT_{\max}$ set to mean and maximum *RT* over all subjects. rt values are significantly lower for sorting unknown objects than for known objects (WSR, $z = 3.94$, $p < 0.001$, $Mdn = 0.83$ (unknown) vs. 1.0 (known)) and for non-sorting advice (WSR, $z = 4.62$, $p < 0.001$, $Mdn = 1.0$ (non-sorting)). In summary, in Ex.2 the calculated rt-indicator can distinguish between certain and uncertain advice. Even tough, a higher uncertainty can not always be equated with incorrect advice, it can avoid following incorrect advice for unknown objects.

In Fig. 6(g-i), we compare *FB* for Ex.1A&B and Ex.2. We excluded 7 bearded subjects since facial hair drastically reduced the quality of extracted AUs. Even though Fig. 6(g) shows a trend of a higher *FB* for sorting advice over non-sorting in Ex.1B, WSR indicates no significant differences. Fig. 6(h) shows *FB* for Ex.2, comparing non-sorting and sorting advice for known and unknown objects. There is a significant difference between sorting known (green, $Mdn = 0.02$) and unknown objects (red, $Mdn = 0.03$), (WSR, $z = 2.49$, $p < 0.05$). In Fig. 6(i), we calculate f (Eq. (5)) with $FB_{\min} = 0.1$ and $FB_{\max}$ set to maximum *FB* over all subjects. For this calculation, most advice would be considered as certain. Results show higher f-values for sorting unknown objects for 9 subjects but do not allow to clearly distinguish certain from uncertain advice. Nevertheless, visual inspection of the recordings suggests a connection between facial dynamics and uncertainty (Fig. 5(d)). A deeper analysis of relevant AUs and refined calculation of *FB* is needed to further evaluate the suitability of facial dynamics as an indicator for incorrect advice. We compare IOP and LOP (Fig. 6(b-c)) as two exemplary methods to combine rt and f to b-scores (Eq. (8)). IOP often misclassified uncertain advice as $b = 1.0$. LOP is more robust and has significantly lower b for sorting unknown compared to known (WSR, $z = 3.11$, $p < 0.005$, $Mdn = 0.90$ (unknown), 1.0 (known)) and non-sorting (WSR, $z = 4.01$, $p < 0.001$, $Mdn = 1.0$ (non-sorting)).

V. Conclusion and Outlook

We introduced LUNAA, a new algorithm to learn from unreliable action advice in IRL. In contrast to related approaches, we use a state-dependent trust in human action

advice based on three indicators, i.e. consistency, retrospective optimality, and behavioral cues. Evaluations in a gridworld scenario and a robotic sorting task showed that for partially incorrect advice LUNAA outperforms a state-independent computation of trust in human advice as proposed in related works. In the gridworld setting, results demonstrate that behavioral cues can be particularly useful in the case of consistent wrong advice. Therefore, we additionally evaluated response times and facial dynamics as two examples of behavioral cues in three robotic tasks with 28 participants. In these experiments, response times allowed to clearly distinguish between certain and uncertain human advice. Facial dynamics showed promising first results, however need further investigation. In future work, we plan to investigate occurrences of behavioral cues with more participants in a larger variety of tasks to develop a deeper understanding of correlations between human behavioral uncertainty and partially wrong advice and explore differences between individuals. Here, we consider verbal articulation [26] as a promising additional signal. Moreover, we plan to explore how to replace the currently state-independent self-confidence in the agent's Q-function by a state-dependent approach, e.g. using Bayesian RL. Finally, using the indicators to help a robot understand underlying reasons for incorrectness of advice and learn a model of the teacher's capabilities, is another interesting direction.

REFERENCES

[1] A. L. Thomaz, G. Hoffman, and C. Breazeal, "Real-time interactive reinforcement learning for robots," in *AAAI 2005 Workshop on Human Comprehensible Machine Learning*, 2005.

[2] W. B. Knox and P. Stone, "Tamer: Training an agent manually via evaluative reinforcement," in *2008 7th IEEE International Conference on Development and Learning*. IEEE, 2008, pp. 292–297.

[3] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[4] G. Li, R. Gomez, K. Nakamura, and B. He, "Human-centered reinforcement learning: A survey," *IEEE Transactions on Human-Machine Systems*, vol. 49, no. 4, pp. 337–349, 2019.

[5] T. A. Kessler Faulkner and A. Thomaz, "Interactive reinforcement learning from imperfect teachers," in *Companion of the 2021 ACM/IEEE international conference on human-robot interaction*, 2021, pp. 577–579.

[6] D. Koert, M. Kircher, V. Salikutluk, C. D'Eramo, and J. Peters, "Multi-Channel Interactive Reinforcement Learning for Sequential Tasks," *Frontiers in Robotics and AI*, vol. 7, no. September, pp. 1–19, 2020.

[7] S. Griffith, K. Subramanian, J. Scholz, C. L. Isbell, and A. L. Thomaz, "Policy shaping: Integrating human feedback with reinforcement learning," *Advances in neural information processing systems*, 2013.

[8] W. B. Knox and P. Stone, "Reinforcement learning from simultaneous human and mdp reward." in *AAMAS*, 2012, pp. 475–482.

[9] T. A. Kessler Faulkner, E. Schaertl Short, and A. L. Thomaz, "Interactive Reinforcement Learning with Inaccurate Feedback," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 7498–7504, 2020.

[10] A. L. Thomaz, C. Breazeal *et al.*, "Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance," in *Aaai*, vol. 6. Boston, MA, 2006.

[11] Z. Lončarević, A. Ude, B. Nemec, A. Gams *et al.*, "User feedback in latent space robotic skill learning," in *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2018, pp. 270–276.

[12] A.-L. Vollmer and N. J. Hemion, "A user study on robot skill learning without a cost function: Optimization of dynamic movement primitives via naive user feedback," *Frontiers in Robotics and AI*, 2018.

[13] A. Najar and M. Chetouani, "Reinforcement learning with human advice: a survey," *Frontiers in Robotics and AI*, vol. 8, 2021.

[14] R. Arakawa, S. Kobayashi, Y. Unno, Y. Tsuboi, and S.-i. Maeda, "Dqn-tamer: Human-in-the-loop reinforcement learning with intractable feedback," *arXiv preprint arXiv:1810.11748*, 2018.

[15] F. Cruz, J. Twiefel, S. Magg, C. Weber, and S. Wermter, "Interactive reinforcement learning through speech guidance in a domestic scenario," in *2015 international joint conference on neural networks (IJCNN)*. IEEE, 2015, pp. 1–8.

[16] A. Yazidi, B. J. Oommen, and M. Goodwin, "On distinguishing between reliable and unreliable sensors without a knowledge of the ground truth," in *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, vol. 2. IEEE, 2015, pp. 104–111.

[17] K. Ni and G. Pottie, "Bayesian selection of non-faulty sensors," in *2007 IEEE International Symposium on Information Theory*. IEEE, 2007, pp. 616–620.

[18] R. Loftin, B. Peng, J. MacGlashan, M. L. Littman, M. E. Taylor, J. Huang, and D. L. Roberts, "Learning behaviors via human-delivered discrete feedback: modeling implicit feedback strategies to speed up learning," *Autonomous agents and multi-agent systems*, 2016.

[19] M. E. Taylor and A. Borealis, "Improving reinforcement learning with human input." in *IJCAI*, vol. 328, 2018, pp. 5724–5728.

[20] Z. Lin, B. Harrison, A. Keech, and M. O. Riedl, "Explore, exploit or listen: Combining human feedback and policy model to speed up deep reinforcement learning in 3d worlds," *arXiv preprint arXiv:1709.03969*, 2017.

[21] A. Kurenkov, A. Mandlekar, R. Martin-Martin, S. Savarese, and A. Garg, "Ac-teach: A bayesian actor-critic method for policy learning with an ensemble of suboptimal teachers," *arXiv preprint arXiv:1909.04121*, 2019.

[22] M. Sridharan, "Augmented reinforcement learning for interaction with non-expert humans in agent domains," in *2011 10th International Conference on Machine Learning and Applications and Workshops*, vol. 1. IEEE, 2011, pp. 424–429.

[23] T. Love, R. Ajoodha, and B. Rosman, "Should i trust you? incorporating unreliable expert advice in human-agent interaction."

[24] W. B. Knox and P. Stone, "Combining manual feedback with subsequent mdp reward signals for reinforcement learning," in *Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1*. Citeseer, 2010, pp. 5–12.

[25] X. He, H. Chen, and B. An, "Learning behaviors with uncertain human feedback," in *Conference on Uncertainty in Artificial Intelligence*. PMLR, 2020, pp. 131–140.

[26] E. Krahmer and M. Swerts, "How children and adults produce and perceive uncertainty in audiovisual speech," *Language and speech*, vol. 48, no. 1, pp. 29–53, 2005.

[27] A. R. Bland *et al.*, "Different varieties of uncertainty in human decision-making," *Frontiers in neuroscience*, vol. 6, p. 85, 2012.

[28] M. Greis, J. Karolus, H. Schuff, P. W. Wozniak, and N. Henze, "Detecting uncertain input using physiological sensing and behavioral measurements," *ACM International Conference*, pp. 299–304, 2017.

[29] A. E. Hramov, N. S. Frolov, V. A. Maksimenko, V. V. Makarov, A. A. Koronovskii, J. Garcia-Prieto, L. F. Antón-Toro, F. Maestú, and A. N. Pisarchik, "Artificial neural network detects human uncertainty," *Chaos: An Interdisciplinary Journal of Nonlinear Science 28.3, 033607*, 2018.

[30] S. Lallé, C. Conati, and G. Carenini, "Predicting confusion in information visualization from eye tracking and interaction data," *IJCAI International Joint Conference on Artificial Intelligence*, vol. 2016-January, pp. 2529–2535, 2016.

[31] J. Gonzalez-Billandon, A. M. Aroyo, A. Tonelli, D. Pasquali, A. Sciutti, M. Gori, G. Sandini, and F. Rea, "Can a robot catch you lying? a machine learning system to detect lies during interactions," *Frontiers in Robotics and AI*, p. 64, 2019.

[32] Y. Cui, Q. Zhang, A. Allievi, P. Stone, S. Niekum, and W. B. Knox, "The empathic framework for task learning from implicit human feedback," *Conference on Robot Learning*, 2020.

[33] C. Arzate Cruz and T. Igarashi, "A survey on interactive reinforcement learning: Design principles and open challenges," in *Proceedings of the 2020 ACM designing interactive systems conference*, 2020.

[34] Z. Wang and M. E. Taylor, "Improving reinforcement learning with confidence-based demonstrations." in *IJCAI*, 2017, pp. 3027–3033.

[35] B. Amos, B. Ludwiczuk, and M. Satyanarayanan, "Openface: A general-purpose face recognition library with mobile applications," Tech. Rep., 2016.

[36] P. Ekman and W. V. Friesen, "Facial action coding system," *Environmental Psychology & Nonverbal Behavior*, 1978.

[37] M. Stone, "The linear opinion pool," *Ann. Math. Statist*, vol. 32, pp. 1339–1342, 1961.

[38] J. O. Berger, "Statistical decision theory," in *Game Theory*. Springer, 1989, pp. 217–224.