Yesterday's Reward is Today's Punishment: Contrast Effects in Human Feedback to Reinforcement Learning Agents

Divya Ramesh
University of Michigan - Ann Arbor
Ann Arbor, MI
dramesh@umich.edu

Anthony Z. Liu
University of Michigan - Ann Arbor
Ann Arbor, MI
anthliu@umich.edu

Andres J. Echeverria
University of Michigan - Ann Arbor
Ann Arbor, MI
andreseg@umich.edu

Jean Y. Song
University of Michigan - Ann Arbor
Ann Arbor, MI
jyskwon@umich.edu

Nicholas R. Waytowich
US Army Research Laboratory
Aberdeen, MD
nicholas.r.waytowich.civ@mail.mil

Walter S. Lasecki
University of Michigan - Ann Arbor
Ann Arbor, MI
wlasecki@umich.edu

ABSTRACT

Autonomous agents promise users of a personalized future, allowing them to direct their attention to tasks most meaningful to them. However, the demands of personalization stand unfulfilled by current agent training paradigms such as machine learning, which require many orders of data to train agents on a single task. In sequential decision making domains, Reinforcement Learning (RL) enables this need, when a priori training of desired behaviors is intractable. Prior work has leveraged user input to train agents by mapping them to numerical reward signals. However, recent approaches have identified inconsistent human feedback as a bottleneck to achieving best-case performance. In this work, we present empirical evidence to show that human perception affected by contrast effects distorts their feedback to Reinforcement Learning agents. Through a set of studies involving 900 participants from Amazon Mechanical Turk who were asked to give feedback to RL agents, we show that participants significantly underrate an agent's actions after being exposed to an agent of higher competence on the same task. To understand the significance of this effect on agent performance during training, we then simulate trainers that underrate actions of an agent based on past performance – creating a systematically skewed feedback signal – integrated into an actor-critic framework. Our results show that agent performance is reduced by up to 98% in the presence of systematic skews in human feedback in Atari environments. Our work provides a conceptual understanding of a source of inconsistency in human feedback, thus informing the design of human-agent interactions.

KEYWORDS

Human-Agent Interaction, Contrast Effects, Reinforcement Learning

ACM Reference Format:

Divya Ramesh, Anthony Z. Liu, Andres J. Echeverria, Jean Y. Song, Nicholas R. Waytowich, and Walter S. Lasecki. 2020. Yesterday's Reward is Today's Punishment: Contrast Effects in Human Feedback to Reinforcement Learning Agents. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 8 pages.

1 INTRODUCTION

Autonomous agents promise a future where humans direct their attention to only those tasks that are most meaningful to them. The key to this future is hyper-personalization – agents catering to the needs of each user by learning each individual's behavioral preferences. Such a promise suggests that agents will be able to learn from limited records of previous interactions with users, adapt to novel situations, and hold multi-purpose capabilities. However, many of these needs stand unfulfilled by current agent training paradigms such as machine learning and machine teaching, which require many magnitudes of data to train agents on a single task.

Fortunately, in domains that require agents to make sequential decisions, this need can be enabled by Reinforcement Learning (RL) – a technique that trains agents through interactions with their environments [26]. Prior work in RL has also leveraged user input in the form of feedback mapped to numerical reward signals. [1, 6, 12, 16, 23, 29]. For example, in domains like autonomous driving, where the cost of learning by trial-and-error is too expensive [13, 25], there have been attempts to use human drivers' actions as rewards and punishments to design robust self-driving algorithms [14].

Such human-in-the-loop RL methods have gained added traction because of their ability to drastically reduce the large training time requirements traditionally associated with RL algorithms [29]. However, several of these methods have noted inconsistencies such as drift & sparsity in user input over time, and changing preferences which prevent agents from achieving optimal performance [12, 16, 29]. Researchers have also offered several interpretations of human inputs, such as using them as advice, communication or policy feedback, to handle inconsistencies [8, 10, 22, 27].

However, sequential decision-making being a routine aspect of human lives, errors in human decision-making has received considerable attention in behavioral psychology [28]. Studies on sequential tasks in other domains have shown that inconsistencies in human judgement arise from perceptual context errors i.e., contrast effects [15, 18, 30]. For instance, in performance judgements of Olympic gymnastics, it was found that judges often evaluated an athlete's performance based on perceptions of preceding athlete's competence [7]. While evidence of such contextual errors in evaluative feedback has been previously observed in human-agent interaction settings, their existence stands unverified [8, 10, 12, 17, 27].

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

Furthermore, while the presence of such errors is believed to be a bottleneck to achieving maximal agent performance, there exists no quantitative understanding of the degradation in performance caused by their existence. Thus, in an attempt to bridge this gap, we ask and answer the following research questions:

- RQ 1: Does a person's prior perception of an agent's competence influence their subsequent evaluations of the agent's actions?
- RQ 2: What impact does this influence have on the agent's performance when the agent is trained from human feedback?

From a set of six randomized control trials involving 900 participants from Amazon's Mechanical Turk giving feedback to agents playing several Atari games, we present empirical evidence to show that human evaluations of agent actions are influenced by prior perceptions of agent competence on the same task. The influence is found to be contrastive in nature – i.e., *humans significantly underrate an agent actions when previously exposed to an agent of higher competence on the same task* and vice versa.

We then carry out a set of controlled simulation studies on various trainer feedback patterns in four Atari environments to understand the significance of this effect on agent performance during training. Our results show that *in the presence of human feedback subject to contrast effects, agent performance reduces by up to 98% in Atari environments*. This implies that even a gradual but steady decline of feedback signals (i.e., skewed feedback signals) to RL agents results in agents learning policies vastly different from the optimal. Hence, our results call for careful consideration of both - human biases and variance & stability characteristics of algorithms when incorporating human feedback in RL.

In this paper, we provide a conceptual understanding of a source of inconsistency in human feedback to RL agents. To our knowledge, this is the first work that seeks to provide a unifying explanation of a phenomenon observed in human-agent interaction settings and other sequential decision making domains, thus informing the design of human-agent interactions.

2 RELATED WORK

We review the literature on recent strategies devised to incorporate human input in training reinforcement learning agents, challenges faced, and draw parallels with similar sequential decision-making tasks in other domains.

2.1 Human Feedback in Reinforcement Learning

Reinforcement learning (RL) is a technique to solve sequential decision making problems where the agent learns a policy through by sampling actions and maximising rewards associated with the actions [26]. A reward function is thus an integral component of RL, determining the agent's policy and overall behavior. User inputs are often incorporated into the RL framework by augmenting, inferring or replacing such a reward function [6, 8, 11, 12, 17, 29].

There are several different ways to leverage human input in human-in-the-loop RL. For instance, they can be mapped to binary reward signals to train Q-value functions [12, 16, 29], or mapped to advantage functions [2, 23], or to behavioral preferences to infer reward functions which can then be used to train agents [6, 11].

Several challenges have surfaced in incorporating human input. For instance, Isbell Jr and Shelton observed that user rewards reduced exponentially as the agents showed improved performance [12]. Knox and Stone noted that human trainers had a tendency to reward agents positively, creating unwanted feedback loops [17]. Thomaz and Breazeal noted that users changed their reward schemes as they developed mental models of their agents. Several attempts have also been made to explain these inconsistencies. For instance, Thomaz and Breazeal suggest that human feedback may be best interpreted as advice [27]. Griffith et al. suggest treating human feedback as policy-dependent, and Ho et al. have noted that human feedback may be better interpreted as communication rather than rewards and punishments [8, 10].

Our work seeks to provide a consistent explanation for these observed issues of drift & sparsity in human input over time, and seemingly changing user preferences.

2.2 Human Decision Making and Perceptual Context Effects

It is well known in psychology that human judgement and decision making of artifacts is based not only on the inherent value of the artifacts, but also the context surrounding it [28]. Specifically, in the context of sequential decision making, several works have investigated locally occurring errors – where a preceding trial influences the judgement of the current trial. For instance, Hartzmark and Shue have shown that even experienced investors mistakenly perceive earnings news today as more impressive if yesterday's earnings surprise was bad, and less impressive if yesterday's surprise was good, thereby distorting market prices [9]. This type of error where the value of a previously observed signal inversely biases perception of the next signal is termed as *contrast effects*, and has been found in various contexts, including performance judgements in Olympic synchronized swimming [7], batch annotation tasks in crowdsourcing [21], ratings of student's essays [3], judicial decision making [15], and perceptions of partner attractiveness in speed dating settings [5].

Our work attempts to tie the two threads of literature together - to investigate if human feedback to RL agents is subject to contrast effects, and how this effect may affect agent training and hence subsequent feedback from humans in sustained interactions.

3 PROBLEM STATEMENT

In this section we provide preliminaries relevant to the design of our study investigating contrast effects in human-agent interactions.

For our experiments, we assume a Markov Decision Process (MDP) formulation of RL, $(\mathcal{S}, \mathcal{A}, T, R)$, the state space, action space, transition function, reward function respectively. In this formulation, an agent tries to maximize the reward it earns over time $\mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t R(s_t, s_{t+1})\Big]$, for some discount factor γ.

In order to study the interaction effects between human trainers and RL agents, we choose the Atari Arcade Learning Environment [4]. Our choice of the environment is motivated by the following:

- Task Complexity: Atari environments are fairly complex, capturing many of the interactions in the real-world.
- Algorithmic Performance of RL Agents: Many state-of-the-art learning algorithms have shown human-comparable performance on Atari games.
- Human Perceptible Task Outcomes: Most humans are able to easily understand the rules and evaluation criteria of the Atari games.
- Reproducibility: Videos of agents playing games in this environment can be easily recorded and streamed on Amazon Mechanical Turk (our target participant pool for the experimental study).

We use the actor-critic agent, Advantage Actor-Critic (A2C) in our experiments. Actor-critic methods combine an agent, the actor, that produces continuous actions without considerations for value function optimization, and a critic that evaluates performance of the actor to achieve optimal policies with low variance [24]. Since humans act as critics of an agent's policy, it is argued that actor-critic methods are well-suited for human feedback [2, 8, 10, 27]. A2C, and its asynchronous variant (A3C) are state-of-the-art actor-critic methods most suitable for Atari environments [24].

The hyperparameters of A2C algorithm have been tuned for a set of six games, which have also been used in prior studies of human feedback in reinforcement learning [6, 11]. Based on participants' ease of understanding of these six games assessed via an informal pilot study, we use a subset(n=3) – Pong, Beam Rider and Breakout for our experimental study.

For our study, we use the following definition of contrast effects - *an effect where the judgement of the current trial is shifted in a direction away from the preceding trial.* We define a unit of 'rewardable action' (henceforth action) as the sequence of steps carried out by the agent from the initial state until the agent reaches a terminal state. The score obtained by the agent in this process serves as a proxy for competence and performance of the agent. In our settings and experiments, participants watch the agent perform actions in the various games and give a numerical rating as feedback.

4 AN EXPERIMENTAL STUDY OF CONTRAST EFFECTS

In this section, we describe our experimental study testing how participants give feedback to agents playing Atari games, when subject to different prior perceptions of the agent competence.

4.1 Experimental Setup

Subjects. Recruitment of subjects and the experiments were in accordance with our institution's IRB policies. As the study fell under federal exemption category 3(i)(A) and/or 3(i)(B) at 45 CFR 46.104(d)), the IRB application was exempted by our institution. Participants were remote crowd workers located in the United States, aged 18 years or older recruited via Amazon's Mechanical Turk crowdsourcing platform, who were compensated at the rate of $4.5 per task ($11/hr) for their participation. Each participant watched and provided numerical evaluations on the actions of reinforcement learning agents playing Atari games. Each worker could only participate once to avoid any learning bias.

Preparation. Here, we summarize the study design and how we prepared the experiment videos.

Videos: The videos depicted asynchronous actor-critic (A2C) agents playing Atari games, recorded at various times during an agent's training session. Each video snippet contained exactly one game play of the agent, starting from the initial state until it reached a termination state. The videos contained agents in one of three distinct competence levels: low, moderate, or high. The skill level of an agent was classified as low or high depending on whether the agent obtained 10% or 100% of the maximum attainable score by a state-of-the-art Asynchronous Actor Critic agent on the same Atari game. The moderate level agents had scores that were in the 40-60% range of the maximum attainable score. The videos with agents in moderate competence levels were used as the *target videos* in our study. There were two video snippets sampled for each competence level to avoid any potential biases arising from choosing a single video. (A formative study carried out previously in the lab verified that selecting agents by the score reached indeed served as proxy indicators of the skill-level of the agents). In addition, the game score of the agents were hidden from the participants in order to motivate realistic agent training scenarios, where there are fewer objective instruments of comparison available as decision-making aids.

Rating Scale: A 5-point Likert rating system was given to the participants to rate the agent's performance of an action in the videos. It is established that such a 5-point scale can reliably capture moderate, neutral and extreme attitudes of participants [20]. Additionally, we provided textual descriptions i.e., 1-Terrible, 2-Poor, 3-Average, 4-Good, and 5-Excellent to aid a precise and stable understanding of the meaning of each point on the scale, thus preserving the reliability and validity of measurements. We verified participants' understanding and their agreement of rating criteria via a formative study and context manipulation check step (Table 1).

Study Design: The experiment was designed as a between-subject study (participants were in unique prior manipulation conditions). 150 unique workers were recruited per task, and were randomly assigned to one of three experimental conditions.

Experimental Conditions. Below we explain the different experimental conditions we used in our study.

Control Condition: In the control condition, participants watched and rated two videos of an agent that had acquired a moderate level of competency (scoring 50% of the maximum attainable score) in the Atari games. In all our trials, the ratings of the first video in this group served as ground-truth for comparison in our analyses.

Low Condition: Participants assigned to the low condition first watched and rated an agent that had just begun to learn the game before proceeding to watch and rate the target video.

High Condition: Participants first watched an agent that had learned to play the game at the level of an expert human player or better before proceeding to watch and rate the target video.

Procedures. Before beginning the study task, participants first viewed an instruction page that described the task, and asked for their consent to participate. Once they provided consent, each participant was asked two questions that tested their understanding of the task and game in general. We later used the answers provided to this question as an attention check, and filtered out participants with wrong answers for these questions. Each participant then watched and rated the manipulation video, followed by the target video. No explicit evaluation guidelines were given to participants to avoid any potential biases of scores. After this, participants were asked a few questions that to obtain a qualitative understanding of their assessments and their overall experience during the study.

4.2 Results

In this section, we report on the analysis of human-agent interaction effects from three Atari games: Pong, Beam Rider, and Breakout.

Human evaluations of agent actions are subject to perceptual context effects. We show this using two steps. First, we verified that the agents in the prior context videos were perceived as intended, e.g., verified if the videos of agents with low competence received the lowest rating among all videos, videos of agents with high competence received the highest rating among all the videos, and if ratings for the videos of agents with moderate competence fell in between [19]. Table 1 summarizes the average and standard deviation of ratings of each prior context video. We ran Welch's t-test with Bonferroni correction (a total of three comparisons) to test if the difference of ratings between the level of competences are significant. For all comparisons, the difference was significant ($p < .05$). Therefore, we conclude that there was a successful manipulation of context.

Table 1: Participants' perceptions of agent competence in the three different experimental conditions

		LOW	CONTROL	HIGH
Pong	Average	1.94	2.63	4.50
	Stdev	1.13	0.94	0.75
Beam Rider	Average	2.12	3.03	3.68
	Stdev	0.77	0.80	0.66
Breakout	Average	1.16	4.06	4.50
	Stdev	0.37	0.75	0.72

Second, we verified that the ratings of target videos varied across the different manipulation conditions. A Kruskal-Wallis Test indicated a difference across all conditions in all of the 3 games: Pong ($U = 20.73, p < .0001$), Beamrider ($U = 23.08, p < .0001$), and Breakout ($U = 19.96, p < .0001$).

Human evaluators tend to compare and contrast their evaluations with previously observed agents. Using the control condition as the ground-truth for comparison of ratings for target videos, we observed the following.

In all the three games, participants in the low competence agent condition rated the target agent's actions higher than the control condition (ground-truth) - Pong ($U = 418.0, L = 32, C = 38, and p < .005$), Beam Rider ($U = 465.0, L = 41, C = 34, and p < .01$) and Breakout ($U = 343.0, L = 31, C = 36, and p < .001$).

Additionally, in all the three games, participants in the high competence agent condition rated the target agent's actions lower than the control condition (ground truth) - Pong ($U = 436.0, H = 34, C = 38, and p < .005$), Beam Rider ($U = 499.0, H = 40, C = 34, and p < .0001$), Breakout ($U = 411.5, H = 32, C = 36, and p < .01$).

The ratings for the target videos across all three experimental conditions, by each game is shown in Fig. 1.

From our analysis, the hypothesis that participants' prior perception of agent's competence has a contrastive influence on their subsequent evaluations of agent actions receives strong support.

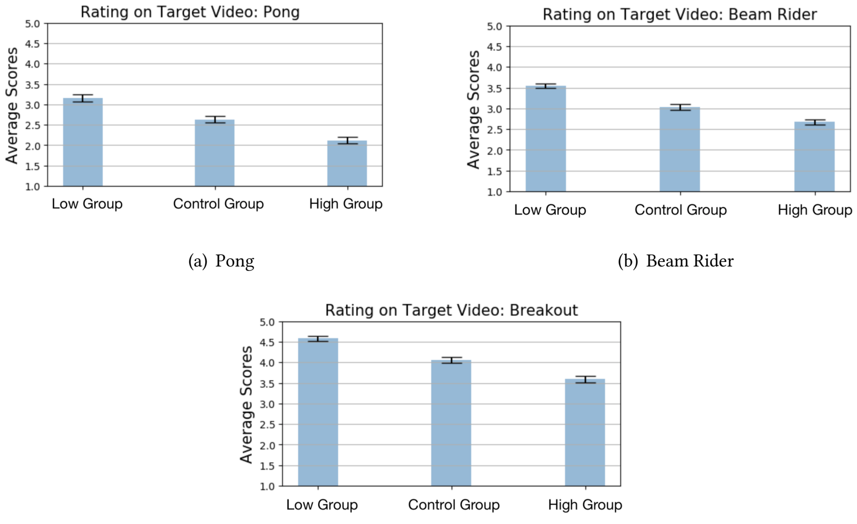
Exploratory Findings. From our formative study, we noted that the agent score was a salient source of information that served as a decision aids in human evaluations of agents. Therefore, we conduct more studies to explore if the presence of an implicit objective standard of comparison, such as the game score, has an influence on the effects.

We choose to look at Pong, the game that participants were most familiar with for further examination. The experiment was run with the the same videos as described earlier, except this time, the agent's game score was made visible to the participants. We observe a context effects. A Kruskal-Wallis test shows significance with comparison across conditions ($U = 25.68, p < .0001$). On comparing the ratings of each group with the control condition, we see a strong contextual effect in the low competence agent condition ($U = 405.0, L = 41, C = 37, and p < .0001$), and a moderate effect in the participants in the high competence agent condition High v/s Control ($U = 588.0, H = 35, C = 37, and p = .0687$) and there was a significant difference in the ratings of target between the low and the high groups, with the target video in the low competence agent context condition receiving higher ratings than that in the high competence agent context condition ($U = 319.5, L = 41, H = 35, and p < .0001$).

We also explore the size of the effect when the target video is sampled at different instances during training. Using the game of Pong, we run studies with the target videos of agents sampled at 40% (4M), 60% (6M) and 80% (8M) of a complete training session. Our preliminary findings indicate that prior context matters at the 40% and 60% levels of agent training, but not at the 80% level of agent training. A Kruskal-Wallis indicates significance at both 40% ($statistic = 20.73, p < .0001$) and 60% ($statistic = 15.33, p < 0.001$) levels. In addition, a comparison of the ratings with control indicates a contrast effect in the 40% and 60% level of agent training.

- 40% training:
  - Low v/s Control ($U = 418.0, L = 32, C = 38$, and $p < .005$)
  - High v/s Control ($U = 436.0, H = 34, C = 38$, and $p < .005$)
- 60% training:
  - Low v/s Control ($U = 570.0, L = 38, C = 40$, and $p < .01$)
  - High v/s Control ($U = 494.0, H = 34, C = 40$, and $p < .01$)
- 80% training:
  - Low v/s Control ($U = 672.5, L = 39, C = 39$, and $p = 0.054$)
  - High v/s Control ($U = 644.5, H = 37, C = 39$, and $p < 0.063$)

We hypothesize that the lack of observation of the effect at 80% training levels ($U = 2.86, p = 0.24$) was due to the almost imperceptible difference between the agents in the target and high conditions. Further investigation is required to understand if the null result was due to the absence of the effect or an artifact of the instrument used to measure the effects.



(a) Pong

(b) Beam Rider

(c) Breakout

Figure 1: Comparison of ratings of the target agent's actions across participants in the three groups for the three games. The error bars indicate standard error range.

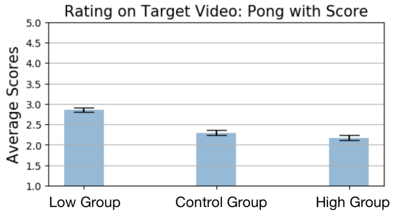


Figure 2: Comparison of ratings of the target agent's actions across participants in the three groups, with the agent's score made visible to participants. The error bars indicate standard error range.

Summary of Results. From our studies, we conclude that,

- Humans overrate an agent's actions when previously exposed to an agent with lower competence on the same task, and conversely,
- Humans underrate an agent's actions when previously exposed to an agent of higher competence on the same task.

5 SIMULATION STUDIES ON HUMAN FEEDBACK PATTERNS UNDER CONTRAST EFFECTS

In this section, we describe the controlled simulation studies conducted in Atari environments to quantify the impacts of contrast effects in human-agent interaction scenarios.

5.1 Trainer Feedback Patterns

Consider an interaction setting where a human trainer is continuously providing feedback to an RL agent learning to play a game of Atari. From experimental results discussed in Section 4, we know that when this trainer provides a sequence of feedback signals to the agents, they are subject to contextual effects. I.e., at any instant of time, their past interactions with the agent influences their subsequent evaluations of agent actions.

Furthermore, in high dimensional state spaces such as Atari games, RL methods experience issues of training instability, where the improvement in agent performance occurs non-monotonically. What this means is that over the course of training, an RL agent may cyclically repeat patterns of high performance in some episodes, immediately followed by lower performance in subsequent episodes – this, until a stable optimal policy is learned. When a trainer is subject to contextual effects, we know from our previous results, that this cyclical pattern of agent behavior induces a contrastive influence on their evaluations. This would mean that for an agent whose average competence is increasing over time, the trainer is more likely to underrate sub-optimal agent actions after witnessing spikes of high performance.

Accounting for this facet of contrast effects, we simulate human feedback such that the reward for the agent actions is a function of the agent's competence at any instant of time.

These feedback patterns are designed to be consistent with the theory of contrast effects [9, 30], and contain several important aspects of real-world biased rewards such as exponential decay [12], damped feedback [10, 17, 22] and varying frequency [17, 29].

5.2 Experimental Setup

We describe how we used the true Atari environment rewards to simulate biased rewards that result from trainers experiencing contrast effects after viewing bursts of high performance of agents.

***Damping Feedback*.** We simulate the human feedback as a function of the agent's competence, measured by the highest episode score achieved by the agent since the start of its training. Thus, the designed function simulates contrast effects as lenient trainer behavior towards the agent when the agent's competence is low, which grows stringent as the agent's competence improves. To simulate this feedback pattern, for each action of the agent, we utilize the environment reward and scale it by a damping function, reflecting the level of leniency or stringency of a simulated trainer at any given time. i.e.,

$$R_h(s_t, a_t) = L(C_t) \cdot R_e(s_t, a_t) \tag{1}$$

where,

$R_h(s_t, a_t)$ is the simulated human reward for a state-action pair (s_t, a_t) at any instant t,

$R_e(s_t, a_t)$ is the original Atari environment reward at the same instant, and

$L(C_t) \in [0, 1]$ is the damping function whose value is initialized to 1 at the start of the training.

$C_t = max(\sum_0^t R_e(s_t, a_t))$ is the competence of the agent at time t.

To account for different possible ways humans are affected by the contrast effects, we test two different damping functions:

- *Linear* – the function is a linearly decaying function over the most recent high score achieved by the agent.
- *Exponential* – the function is an exponentially decaying function over the most recent high score achieved by the agent.

Note that in the game of Pong, there are also negative rewards or penalties. Using the motivation described above, the simulated trainer assigns lower penalty for bad actions when the agent's competence is low, and transitions to giving higher penalty as the agent grows in competence.

***Frequency of Feedback Changes*.** Another important variation to account for in our simulation is the frequency which the trainer's feedback changes by the contrast effect. In our envisioned interactions, it is possible for a trainer to experience contrast effects at different scales. If the trainer's perception of the agent's competence changes with every new high score achieved by the agent, the trainer's feedback patterns may change frequently, and infrequently if otherwise.

- *Micro* – Extremely frequently – the trainer perceives a change in agent competence every time the agent achieves a new highest score,
- *Macro* – Extremely infrequently – the trainer perceives a change in agent competence only when the agent completes 50% of its training.
- *Macro 2M* – In between – the trainer perceives a change in agent competence every 2M timesteps.

***All Simulated Trainer Feedback Conditions*.** Simulating a trainer's feedback incorporates both damping function and frequency of feedback changes, which results in 7 simulation conditions (where one is the baseline condition with ideal rewards, which indicates the absence of contrast effects).

- *Macro + Linear*
- *Macro + Exponential*
- *Macro 2M + Linear*
- *Macro 2M + Exponential*
- *Baseline*
- *Micro + Linear*
- *Micro + Exponential*

5.3 Results

We plot the *actual* reward curves for the four environments and conditions in Figure 3 over 10 million timesteps (while giving the dampened rewards to the agents). To avoid clutter, we plot all conditions except the two Macro 2M conditions. We record the average performance over the last 1 million timesteps of training to gauge the final performance of the agent, which we show in Table 2.

The more frequent the contrast effects, the worse the agent learns. With the simulated trainer's feedback scheme changing every time a new high score was achieved by the agent(micro), all the agents performed poorly (and barely learning anything at all in Pong and Breakout). Under less frequent feedback scheme changes (the baseline and Macro), the performance is minimally affected.

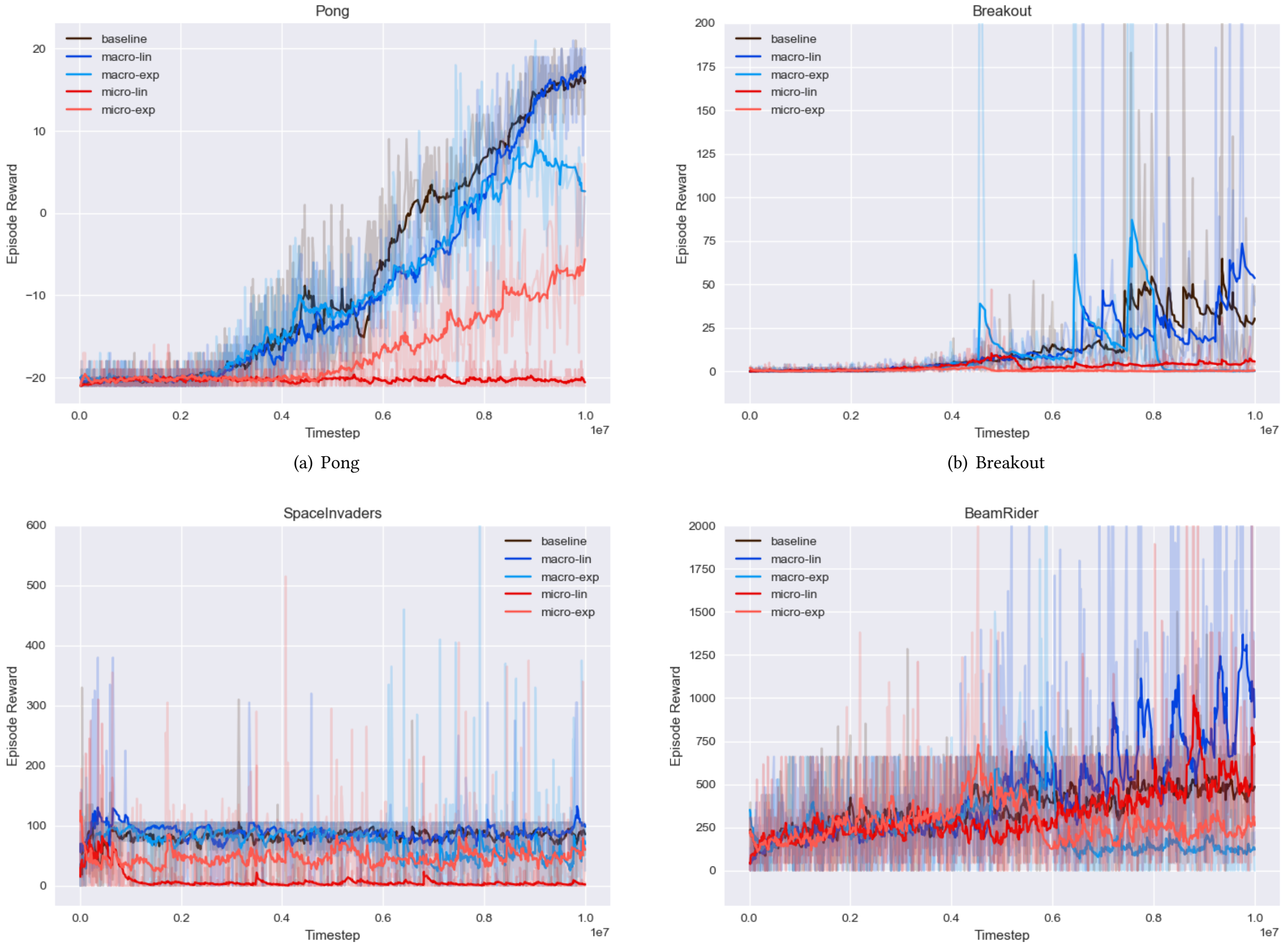
The greater the change in a trainer's feedback, the worse the agent learns. We find that an exponentially changing feedback scheme causes more degradation in performance than a linearly changing feedback scheme. Even in the case of Macro changes, where the performance is minimally affected, a severe drop in the feedback scheme (as in an exponential change) causes more degradation than a gradual drop (as in a linear change) in feedback scheme.

6 DISCUSSION

It is known independently from reinforcement learning and psychology that convergence of agents is affected by the stability of rewards, and that human perception is subject to contrast effects. However, this is the first time that the two concepts are brought together in human-agent scenarios. In doing so, our work shows that human feedback to reinforcement learning agents, previously assumed to be subject to changing human preferences and agent policies [2, 8, 10, 12, 17, 27, 29], is prone to cognitive bias that can be analyzed and modeled in systematic ways. This is a novel contribution to human-agent literature.

As in previously verified settings, contrast effects are strong when the change in agent's policy is perceptible to human trainers. However, unlike other settings, biased feedback to reinforcement learning agents is augmented by changes in the agent's policies. This propagates back to agents, potentially creating unstable feedback loops. Our combined experiments hence call for careful considerations of both - human biases and variance & stability characteristics of algorithms to elicit human feedback.

Noisy human feedback has been computationally handled previously with an intuitive understanding that human feedback is



(a) Pong

(b) Breakout

(c) Space Invaders

(d) Beam Rider

Figure 3: Training curves under each simulated trainer feedback scenario. We plot the baseline agents with *ideal* rewards over 10M timesteps of training, while also showing the agents trained under non-ideal conditions of trainer feedback subject to simulated contrast effects.

	baseline	macro				micro	
		halfway		every 2M			
		lin	exp	lin	exp	lin	exp
Pong	15.4	16.0	4.1	6.9	3.8	-20.3	-6.8
Breakout	24.2	40.9	0.4	9.3	0.4	4.4	0.3
SpaceInvaders	83.1	99.8	57.2	104.1	74.8	3.4	43.8
BeamRider	489.4	988.5	121.9	1107.0	140.4	527.5	239.7

Table 2: Average performance in the last 1M timesteps. We color cells based on relative performance with other conditions within the same game. Green=Best, Red=Worst.

dependent on the agent policy observed by a human trainer [8, 23]. Contrast effects differ from other noise forms(such as stochasticity and adversarial noise) primarily in that they are strong only when the change in agent's policy is perceptible to human trainers. However, it is known that these effects may be attenuated with awareness and experience of human trainers [5, 9, 18, 30].

Contrast effects being predictable, computational approaches such as Bayesian and regression modeling may also reliably model the bias [5, 9]. Future work investigating these effects in continued sequential interactions between humans and agents may provide more insights into appropriate computational solutions.

7 CONCLUSION

Human perception affected by the effects of context distorts their feedback to reinforcement learning agents in continuous interaction settings. In this paper, we experimentally verified that human evaluators subjected to perceptual contrast effects underrate (or overrate) an agent's actions when previously exposed to an agent with higher (or lower) competence on the same task. Furthermore, we show that not accounting for these effects when incorporating human feedback in on-policy reinforcement learning methods leads to deleterious outcomes in agent training procedures. Our work seeks to provide a conceptual framework that inspires the design of feedback mechanisms in human-in-the-loop reinforcement learning systems.

ACKNOWLEDGMENTS

We would like to thank Nikola Banovic, Zhezheng (Chris) Chen, Vaishnav Kameswaran, Jonathan Kummerfeld, Stephanie O'Keefe, Emily Mower Provost, members of the Interactive Systems Lab at the University of Michigan, and the anonymous reviewers for helpful feedback and suggestions. We also thank the anonymous workers from Amazon Mechanical Turk for their participation. This research was supported in part by the United States Army Research Laboratory (ARL). Any opinions, findings, conclusions or recommendations expressed above do not necessarily reflect the views of ARL.

REFERENCES

[1] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35, 4 (2014), 105–120.

[2] Dilip Arumugam, Jun Ki Lee, Sophie Saskin, and Michael L Littman. 2019. Deep reinforcement learning from policy-dependent human feedback. *arXiv preprint arXiv:1902.04257* (2019).

[3] Yigal Attali. 2011. Sequential effects in essay ratings. *Educational and Psychological Measurement* 71, 1 (2011), 68–79.

[4] Marc G Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. 2013. The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research* 47 (2013), 253–279.

[5] Saurabh Bhargava and Ray Fisman. 2014. Contrast effects in sequential decisions: Evidence from speed dating. *Review of Economics and Statistics* 96, 3 (2014), 444–457.

[6] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*. 4299–4307.

[7] Lysann Damisch, Thomas Mussweiler, and Henning Plessner. 2006. Olympic medals as fruits of comparison? Assimilation and contrast in sequential performance judgments. *Journal of Experimental Psychology: Applied* 12, 3 (2006), 166.

[8] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L Isbell, and Andrea L Thomaz. 2013. Policy shaping: Integrating human feedback with reinforcement learning. In *Advances in neural information processing systems*. 2625–2633.

[9] Samuel M Hartzmark and Kelly Shue. 2018. A tough act to follow: Contrast effects in financial markets. *The Journal of Finance* 73, 4 (2018), 1567–1613.

[10] Mark K Ho, Michael L Littman, Fiery Cushman, and Joseph L Austerweil. 2015. Teaching with rewards and punishments: Reinforcement or communication?. In *CogSci*.

[11] Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg, and Dario Amodei. 2018. Reward learning from human preferences and demonstrations in Atari. In *Advances in neural information processing systems*. 8011–8023.

[12] Charles Lee Isbell Jr and Christian R Shelton. 2002. Cobot: A social reinforcement learning agent. In *Advances in neural information processing systems*. 1393–1400.

[13] Nidhi Kalra and Susan M Paddock. 2016. Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? *Transportation Research Part A: Policy and Practice* 94 (2016), 182–193.

[14] Alex Kendall, Jeffrey Hawke, David Janz, Przemyslaw Mazur, Daniele Reda, John-Mark Allen, Vinh-Dieu Lam, Alex Bewley, and Amar Shah. 2019. Learning to drive in a day. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 8248–8254.

[15] José H Kerstholt and Janet L Jackson. 1998. Judicial decision making: Order of evidence presentation and availability of background information. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 12, 5 (1998), 445–454.

[16] W Bradley Knox and Peter Stone. 2009. Interactively shaping agents via human reinforcement: The TAMER framework. In *Proceedings of the fifth international conference on Knowledge capture*. ACM, 9–16.

[17] W Bradley Knox and Peter Stone. 2015. Framing reinforcement learning from human reward: Reward positivity, temporal discounting, episodicity, and performance. *Artificial Intelligence* 225 (2015), 24–50.

[18] Robin SS Kramer. 2017. Sequential effects in Olympic synchronized diving scores. *Royal Society open science* 4, 1 (2017), 160812.

[19] David A Kravitz and William K Balzer. 1992. Context effects in performance appraisal: A methodological critique and empirical study. *Journal of Applied Psychology* 77, 1 (1992), 24.

[20] Jon A Krosnick and Leandre R Fabrigar. 1997. Designing rating scales for effective measurement in surveys. *Survey measurement and process quality* (1997), 141–164.

[21] Walter S Lasecki, Jeffrey M Rzeszotarski, Adam Marcus, and Jeffrey P Bigham. 2015. The effects of sequence and delay on crowd work. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1375–1378.

[22] James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, David Roberts, Matthew E Taylor, and Michael L Littman. 2017. Interactive Learning from Policy-Dependent Human Feedback. *arXiv preprint arXiv:1701.06049* (2017).

[23] James MacGlashan, Mark K Ho, Robert Loftin, Bei Peng, Guan Wang, David L. Roberts, Matthew E. Taylor, and Michael L. Littman. 2017. Interactive Learning from Policy-Dependent Human Feedback. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70 (ICML'17)*. JMLR.org, 2285–2294.

[24] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*. 1928–1937.

[25] Jean Y Song, Stephan J Lemmer, Michael Xieyang Liu, Shiyan Yan, Juho Kim, Jason J Corso, and Walter S Lasecki. 2019. Popup: reconstructing 3D video using particle filtering to aggregate crowd responses. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, 558–569.

[26] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[27] Andrea L Thomaz and Cynthia Breazeal. 2008. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence* 172, 6-7 (2008), 716–737.

[28] Amos Tversky and Daniel Kahneman. 1986. Judgment under uncertainty: Heuristics and biases. *Judgment and decision making: An interdisciplinary reader* (1986), 38–55.

[29] Garrett Warnell, Nicholas Waytowich, Vernon Lawhern, and Peter Stone. 2018. Deep tamer: Interactive agent shaping in high-dimensional state spaces. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[30] Peter Yeates, Paul O'Neill, Karen Mann, and Kevin W Eva. 2013. 'You're certainly relatively competent': assessor bias due to recent experiences. *Medical education* 47, 9 (2013), 910–922.