# Reinforcement Learning based Embodied Agents Modelling Human Users Through Interaction and Multi-Sensory Perception

**Kory Mathewson and Patrick M. Pilarski**

Departments of Computing Science and Medicine
University of Alberta, Edmonton Alberta Canada
[korym, pilarski] @ ualberta.ca

## Abstract

This paper extends recent work in interactive machine learning (IML) focused on effectively incorporating human feedback. We show how control and feedback signals complement each other in systems which model human reward. We demonstrate that simultaneously incorporating human control and feedback signals can improve interactive robotic systems' performance on a self-mirrored movement control task where a RL-agent controlled right arm attempts to match the preprogrammed movement pattern of the left arm. We illustrate the impact of varying human feedback parameters on task performance by investigating the probability of giving feedback on each time step and the likelihood of given feedback being correct. We further illustrate that varying the temporal decay with which the agent incorporates human feedback has a significant impact on task performance. We found that 'smearing' human feedback over time steps improves performance and we show varying the probability of feedback at each time step, and an increased likelihood of those feedbacks being 'correct' can impact agent performance. We conclude that understanding latent variables in human feedback is crucial for learning algorithms acting in human-machine interaction domains.

## Introduction

Reinforcement learning (RL) agents can learn optimal actions through building models of environments through perceptive sensors during repeated interactions. Often RL agents cooperate interactively with human trainers to solve difficult tasks. Human teachers are a unique component of the environment who may deliver control signals and contextual information through feedback. As human-robot interaction becomes more complex, due to rapid advancements in actuator and sensor technology, a significant *gap* emerges between the number of possible control signals a human can provide and the number of controllable actuators a robotic system. There is often a limited set of control signals which a human can provide, and a large number of robotic system controllable functions.

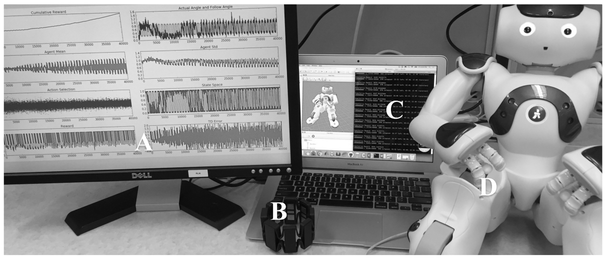

Figure 1. Configuration with A) results example, B) *Myo*, C) Simulation/Learning/Feedback System, and D) *Nao*.

The limit of human provided control signals is of particular interest in the field of robotic prostheses—artificial limbs attached to the body to augment and/or replace abilities lost through injury or illness. Prosthetic limbs with many degrees-of-freedom have been developed [Castellini et al., 2014]. State-of-the-art prosthetics can perform complex functions and movements, but rapid, reactive control of this functionality, by human users, is limited; this limitation causes some users to abandon their devices [Castellini et al., 2014, Biddis et al., 2007, Scheme and Englehart 2011, Micera et al., 2010]. New methods are in development to help humans control complex robotic devices through intelligent control sharing and by allowing a learning agent inside the prosthetic to model the human user. The work presented herein explores RL agents controlled by simulated human electromyography (EMG) signals, with additional reward feedback signals.

## Background

RL is a learning framework inspired by behaviorism [Skinner, 1938] which describes how agents improve over time by taking actions in an environment with a goal of maximizing *expected return*—defined as the cumulative future reward signal received by the agent [Sutton and Barto, 1998]. An agent's control policy is iteratively improved by selecting actions which maximize *return.* RL problems are often described as sequential decision making problems modelled as Markov Decision Processes (MDPs) which define tuples: $(State, Action, Transitions, \gamma, Reward)$, full details of MDPs are omitted for space and can be found in [Sutton and Barto, 1998 and Mathewson *et al.*, 2016]. The ultimate goal of an RL agent is to determine a policy which maps a given current state to the correct actions to maximize *expected return*. In this work we use a continuous actor-critic (AC) algorithm (Algorithm 1) similar to that described in [Pilarski et al. 2011, Pilarski et al. 2013, Mathewson et al. 2016]. AC methods can reduce variance in gradient estimation through the use of two learning systems: a policy-focused actor (selects the best action) and a critic (estimate of value function, criticizes actor) [Sutton and Barto, 1998].

The *Interactive Shaping Problem (ISP)* defines the problem of optimizing the incorporation of human feedback into a learning agent in a sequential decision making problem [Knox and Stone 2010]. The *ISP* asks: how can the agent learn the best possible task policy as measured by task performance or cumulative human feedback, given the information contained in the human feedback [Knox and Stone, 2009 and 2012]. While there are many ways to incorporate human knowledge into a learning system, before and during learning [Thomaz and Breazeal, 2008; Chernova and Tomaz 2014], this paper focuses on incorporating human feedback directly alongside MDP derived reward.

This work builds on the work of Vien and Ertel, who showed that the human feedback model can be generalized to address the problems associated with periods of noisy, and/or inconsistent, human feedback [Vien and Ertel, 2013]. Recent advancements in modelling human feedback with a Bayesian approach have improved on the work of Knox and Stone in discrete environments [Loftin *et al.*, 2015]. Most recently work by MacGlashan *et al.* show that human feedback may be better modelled as an advantage function to handle changes in a human's feedback strategy over time [MacGlashan *et al.*, 2016].

In this study, we explore the implications of varying several latent variables in human feedback for learning algorithms acting in complex human-machine interaction domains. We investigate the probability of the human trainer providing feedback, the probability that feedback is correct, and the effect of *smearing* that feedback over time to account for the limited number of time steps with direct human feedback.

**Algorithm 1** Continuous Actor-Critic RL Algorithm

```
**initialize:** $\mathbf{w}_\mu, \mathbf{w}_\sigma, \mathbf{v}, \mathbf{e}_\mu, \mathbf{e}_\sigma, \mathbf{e}_\mathbf{v}, s$
**repeat:**
    $\mu \leftarrow \mathbf{w}_\mu^T \mathbf{x}(s)$
    $\sigma \leftarrow \exp[\mathbf{w}_\sigma^T \mathbf{x}(s)]$
    $a \leftarrow \mathcal{N}(\mu, \sigma^2)$
    **take action** $a$, **observe** $r, s'$
    $\delta \leftarrow r + \gamma \mathbf{v}^T \mathbf{x}(s') - \mathbf{v}^T \mathbf{x}(s)$
    $\mathbf{e}_\mathbf{v} \leftarrow min[1, \lambda_\mathbf{v} \gamma \mathbf{e}_\mathbf{v} + \mathbf{x}(s)]$
    $\mathbf{v} \leftarrow \mathbf{v} + \alpha_\mathbf{v} \delta \mathbf{e}_\mathbf{v}$
    $\mathbf{e}_\mu \leftarrow \lambda_\mathbf{w} \mathbf{e}_\mu + (a - \mu)\mathbf{x}(s)$
    $\mathbf{w}_\mu \leftarrow \mathbf{w}_\mu + \alpha_\mu \delta \mathbf{e}_\mu$
    $\mathbf{e}_\sigma \leftarrow \lambda_\mathbf{w} \mathbf{e}_\sigma + [(a - \mu)^2 - \sigma^2]\mathbf{x}(s)$
    $\mathbf{w}_\sigma \leftarrow \mathbf{w}_\sigma + \alpha_\sigma \delta \mathbf{e}_\sigma$
    $s \leftarrow s'$
```

# Methods

## Aldebaran Nao and Myo EMG Data

The experimental set up is shown in Figure 1. It is composed of the Aldebaran Nao robotic platform (Aldebaran Robotics), a wireless Myo EMG armband (Thalmic Labs), and a MacBook Air (Apple, 2.2 GHz Intel Core i7, 8GB RAM) for human feedback and running the learning agent.

The experiments in this paper are performed using a simulated Nao platform, a simulated EMG signal, and a simulated human feedback model. We have previously shown the performance of this experimental set-up to be comparable between simulation and real-world experiments [Mathewson *et al.*, 2016]. By simulating the human feedback, we are able to characterize and vary important latent variables hidden from the agent which impact the learning of the system. For this study, we investigate: the rate at which a human-delivered feedback should decay ($smear$), the probability with which the human will provide a feedback ($P(feedback)$), and the probability that this feedback will be correct for a given MDP ($P(correct)$). These are critical variables that have been estimated in previous experiments [Knox and Stone, 2015, Loftin *et al.*, 2015], we aim to improve understanding of their impact through an experimental grid sweep over the variables of interest and investigation into the results.

# Experiments

We extend on the results in [Mathewson *et al.*, 2016] by exploring the impacts of varying model parameters of human trainer feedback on the RL system during the performance of a self-mirrored movement control task. In this task, we preprogram the left arm of the Nao to move in a periodic pattern of flexion and extension at the elbow joint. The RL agent controls the right arm and selects angular displacement actions attempting to match the pattern of the left. With this configuration we are able to define an optimal policy, which would track the pre-programmed arm exactly, with this optimal trajectory we are able to derive MDP reward given a set angular error threshold. When the RL-controlled elbow joint is within the angular deviation threshold of the preprogrammed elbow joint then a reward of 1 is received from the MDP, otherwise, a negative relative error is delivered proportional to the difference between the actual and optimal angles.

We are interested in modelling *smear*, the time-decay with which the feedback given by the human should be decayed. As the human is unable to give feedback at every step that an agent takes, we need to account for the fact that after the exact time step a feedback is given there are likely suboptimal states which support the optimal trajectory. With a decay parameter we are able to smear the human feedback forward in time, it has been shown that the limited human feedback can be applied across near-optimal state-action pairs, and support the agent learning an optimal solution [Pilarski *et al.*, 2011]. We further explore the following characteristics of human feedback: $(P(feedback))$ the probability of giving feedback on each time step, and $(P(correct))$ the probability of giving correct vs. incorrect feedback. These are important latent human parameters to understand, cognitively they represent how effective and attentive a human trainer is.

The continuous state space is defined by the filtered and time averaged dimensionally reduced EMG signal and the angle of the actuated joint, and is represented with approximation using tile coding identically to Mathewson *et al.* [Pilarski *et al.*, 2011 and 2013, Mathewson *et al.,* 2016]. Parameters were set as follows: $\alpha_v = 0.1/m$, $\alpha_\mu = \alpha_\sigma$, $\gamma = 0.9$, $\lambda_w = 0.3$, $\lambda_v = 0.7$, joint angles were limited by manufacturer specifications at $\theta \in [0.0349, 1.5446]$ rads. Weight vectors $\boldsymbol{e}_v$, $\boldsymbol{e}_\mu$, $\boldsymbol{e}_\sigma$, **v**, $\boldsymbol{w}_\mu$ and $\boldsymbol{w}_\sigma$ were initialized to 0 and standard deviation was bounded by $\sigma \geq 0.01$. The eligibility trace update for the critic is scaled by $\gamma$ to speed up convergence. Maximum number of time steps = 10k, learning update and action selection occurred at ~33 Hz or every ~30 ms, and angular deviation threshold was set to $\Delta\theta_{max} = 0.1$, absolute angular joint updates were clipped to 0.1 and actions were selected and performed on every time step.

The ACRL system was trained online with simulated human feedback and simulated EMG control signals (designed to mimic acceptable control signals). Human feedback is integrated into the learning algorithm as reward accumulated on Step 6 of Algorithm 1. Performance was measured by taking the average mean absolute angular error from the last 5k steps. This was done to compare the experimental results after some learning and helped to reduce noise intrinsic in early learning.

This paper presents results of a parameterized grid sweep over three parameters with given estimates of reasonable values: *smear* = (0.2, 0.5, 0.9), *P(feedback)* = (0.03, 0.05, 0.09), *P(correct)* = (0.6, 0.75, 0.9). Additionally, as a control case, n=60 trials without human-feedback were performed. On all time steps MDP reward and human reward were directly summed and applied to the learning agent update (Algorithm 1).

## Results

The results are presented in Figure 2. Results are presented which show performance over a variety of combinations of parameters for the latent variables of interest: *P(feedback)*, *P(correct)* and *smear*. Results indicate that human interaction improves agent performance on a self-mirroring movement task where performance is measured by the mean angular error over the last 5k time steps. Fig. 2A shows that a lower probability of potentially incorrect feedback provides better performance. Fig. 2B shows that there may not be a significant difference in performance when varying the probability of the correctness human feedback, given tested values of *P(feedback)*. This may also be due to the tested values, which were all greater than a 50% chance of being correct. Fig. 2C shows that there is a benefit to selecting a smear decay value appropriate for the task and robotic control system, this parameter may vary task to task and care must be taken when selecting the smear constant. The results indicate that there is benefit to be gained by correct modelling the latent variables associated with human reward signal to allow for true simultaneous incorporation of human control and feedback. These results indicate that the ACRL algorithm robust to a small amount of incorrect feedback.

On average without human-feedback the RL agent was able to attain a mean absolute error on the final 5k steps of 0.22 ± 0.02 (mean ± SEM, n=60). In comparison, the optimal set of parameters (*P(feedback)=0.06, P(correct)=0.6, smear=0.5*) was able to attain a performance of 0.12 ± 0.01 (mean ± SEM, n=7), the worst performing set of parameters (*P(feedback)=0.09, P(correct)=0.9, smear=0.9*) attained a performance of 0.38 ± 0.18 (mean ± SEM, n=4). A total of 232 trials were run over parameter combinations.

## Discussion

The experiments in this paper are performed using a simulated Nao, simulated EMG signal and simulated human feedback. It has been previously shown the performance of this experimental set-up to be comparable between simulation and real-world experiments [Mathewson *et al.,* 2016]. In this related work we explore the degree to which the learning system is affected when incorporating real human feedback. While working in simulation allows rapid iteration and enables testing of many different algorithmic characteristics, simulation is often an easier learning problem than the real-world, due to simplified physics and reduced noise. Future work will address robust modelling real human feedback, and quantify impact of varying feedback density and correctness. We have shown that smearing of human feedback impacts learning, future work will investigate if the decay of human delivered rewards is best modelled as time dependent over task performance and if optimal decay parameters may be learned online.

In this paper we found that modelling the delivery of human feedback can significantly impact the performance of an ACRL algorithm. While we have not optimized for the human feedback characteristics, these results indicate that some human reward paradigms may be preferable to others [Loftin *et al.*, 2015]. This idea is explored in [MacGlashan *et al.,* 2016] where modelling the user feedback as an advantage function, we can understand positive feed back as 'yes, that was good' and negative feedback as 'no, that was bad'. A greater understanding of human reward strategies is required. Personalized robotics will demand perception of human strategies to learn optimal in a very few sample. Future work will focus predicting and optimizing for known and uncertain feedback strategies.

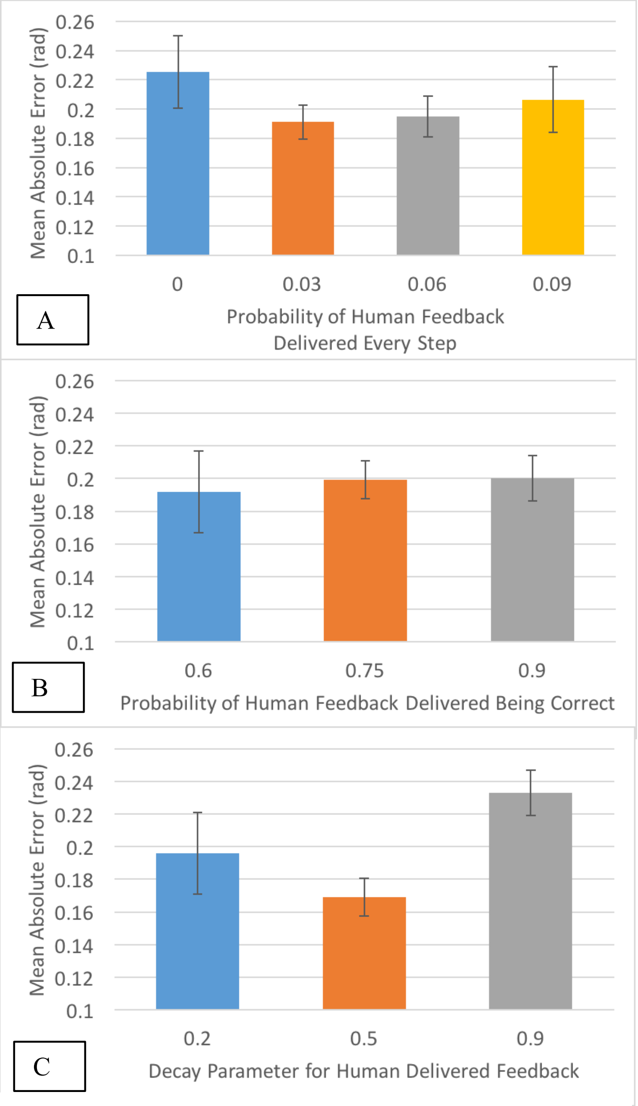

Figure 2. Mean and standard error over experimental conditions A) P(feedback), B) P(correct), C) smear.

Linking control signals in state space with feedback shaping reward signals effectively blends multisensory human data to the learning agent. There remains an open problem of how feedback should best be interpreted by the learning agent and how to encourage human feedback without causing prohibitive additional cognitive load. Modelling, and predicting, human feedback may relieve burden while allowing for shaping control signal interpretation. Human feedback is beneficial to the agent, providing it adds complimentary information about the contextual state the agent is in. Human feedback may shape the MDP reward with more specificity and more often than the sparse, delayed, MDP-derived reward.

Our results demonstrate potential benefits by introducing well modelled human feedback into the robotic learning system. The inclusion of human shaping signals was shown to improve performance over strictly environmentally derived reward. Providing consistent, correct feedback demands cognitive attention from the user which may be difficult if the user is also required to provide control signals to the robotic system. Future work is introduced to explore implications of inviting humans to simultaneously provide control and feedback signals to learning systems.

## Conclusions

This paper contributes a set of results from experiments incorporating simulated human feedback and simultaneous human control in the training of a semi-autonomous robotic agent. These results indicate that task performance increases with the incorporation of human feedback into existing actor-critic reinforcement learning algorithms. These results support the idea that human interaction can improve performance in complex robotic tasks when the human feedback is delivered correctly, consistently, and on a time scale consistent with the original learning problem.

This work supports an emerging viewpoint surrounding human training of a robotic system tightly coupled to a user. By showing improving the performance of the RL agent this work further supports the sharing of autonomy between human and machine.

## Acknowledgements

This work is supported by National Sciences & Engineering Research Council of Canada (NSERC), Alberta Innovates – Technology Futures (AITF), the Canada Research Chairs Program, the Canada Foundation for Innovation, and the Alberta Machine Intelligence Institute (Amii)/Alberta Innovates Centre for Machine Learning (AICML).

# References

[Biddiss and Chau, 2007] Biddiss EA, Chau TT. Upper limb prosthesis use and abandonment: a survey of the last 25 years. *J Prosthet Orthot Intl*. 2007. 31(3):236-57.

[Castellini et al., 2014] Castellini C, Artemiadis P, Wininger M, Ajoudani A, Alimusaj M, Bicchi A, Caputo B, Craelius W, Dosen S, Englehart K, Farina D. *Proc. Of 1st Workshop on Peripheral Machine Interfaces*. 2014.

[Chernova and Tomaz 2014] Chernova S, Thomaz AL. Robot learning from human teachers. *Lecs. AIML*. 2014. 8(3):1-21.

[Knox and Stone, 2009] Knox WB, Stone P. Interactively shaping agents via human reinforcement: The TAMER framework. *In Proc of 5th Intl. Conf. on Knowledge Capture*. 2009 (9-16).

[Knox and Stone 2010] Knox WB, Stone P. 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 2010 May 10 (pp. 5-12). International Foundation for Autonomous Agents and Multiagent Systems.

[Knox and Stone, 2012] Knox WB, Stone P. Reinforcement learning from human reward: Discounting in episodic tasks. *IEEE RO-MAN*. 2012. (pp. 878-885).

[Knox and Stone, 2015] Knox WB, Stone P. Framing reinforcement learning from human reward: Reward positivity, temporal discounting, episodicity, and performance. *AI*. 2015. 225:24-50.

[Loftin et al., 2015] Loftin R, Peng B, MacGlashan J, Littman ML, Taylor ME, Huang J, Roberts DL. Learning behaviors via human-delivered discrete feedback: modeling implicit feedback strategies to speed up learning. *AAMAS*. 2016. 30(1):30-59.

[MacGlashan *et al.*, 2016] MacGlashan J, Littman ML, Roberts DL, Loftin R, Peng B, Taylor ME. Convergent Actor Critic by Humans. 2016. Intl. Conf. on Intelligent Robots and Systems

[Mathewson *et al.*, 2016] Mathewson K, Pilarski PM. Simultaneous control and human feedback in the training of robotic agent with actor-critic reinforcement training. *IJCAI - Interactive Machine Learning Workshop*. 2016.

[Micera et al., 2010] Micera S, Carpaneto J, Raspopovic S. Control of hand prostheses using peripheral information. *IEEE Reviews in Biomedical Engineering*. 2010. *3*, 48-68.

[Pilarski et al., 2011] Pilarski PM, Dawson MR, Degris T, Fahimi F, Carey JP, Sutton RS. Online human training of a myoelectric prosthesis controller via actor-critic reinforcement learning. *IEEE Intl. Conf. on Rehabilitation Robotics*. 2011. (pp. 1-7).

[Pilarski et al., 2013] Pilarski PM, Dick TB, Sutton RS. Real-time prediction learning for the simultaneous actuation of multiple prosthetic joints. *IEEE IC. Rehab. Rob*. 2013. (pp. 1-8). IEEE.

[Scheme and Englehart, 2011] Scheme E, Englehart K. Electromyogram pattern recognition for control of powered upper-limb prostheses: State of the art and challenges for clinical use. *J. of Rehab Res and Dev. 2011*. 48.6: 643.

[Skinner, 1938] Skinner BF. The behavior of organisms: an experimental analysis. *Appleton-Century, Oxford*. 1938.

[Sutton and Barto, 1998] Sutton RS, Barto AG. Reinforcement learning: An introduction. *MIT Press*. 1998.

[Thomaz and Breazeal, 2008] Thomaz AL, Breazeal C. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence*. 2008. 172(6):716-37.

[Vien and Ertel, 2013] Vien NA, Ertel W, Chung TC. Learning via human feedback in continuous state and action spaces. *Applied Intelligence*. 2013. 39(2):267-78.