Characterizing Task Relevant Human Behavior Using a Model Free Metric

Michael Lewis[1], *Lifetime Member*, Katia Sycara[2], *Fellow*, Dana Hughes[2], *Member*, Huao Li[1], *Student Member*, and Tianwei Ni[2], *Student Member*

***Index Terms*—Human-robot teams, Adaptive Systems, Team coordination, Team performance.**

I. Introduction

Transforming human-robot interactions into a common representation is a fundamental problem for HRI. Within psychology and economics normative/descriptive models of humans which compare what they should do (normative) with what they do (descriptive) are common. These models posit true or causally determined behavior that is perturbed by error, for instance, decomposing manual control into a describing function and its remnant. These types of models have not been generally available within HRI because they require both a fixed intent (such as keeping a target centered in the crosshairs) and a well behaved error (such as a Gaussian around a predicted value). When the model to be fit is an uncertain Markov decision process in which actions may be a mixture of the appropriate and extraneous [1] conclude the best approach is a deterministic, worst case evaluation.

We propose a wholly different approach to the problem of disentangling relevant from irrelevant human behavior in HRI. Following [2], [3] we consider human behavior to depend on latent states (such as frustration, trust, attention) in addition to the observable states of a task. As the latent factors are unrelated to the task they can contaminate estimates of a human's policy. We propose to use reinforcement learning to learn a wide diverse set of task-satisfying policies (policy library) any of which might provide a kernel with which extraneous behaviors could be intermixed in producing observed behavior. By matching human trajectories with these task-satisfying policies we can find the policy that best reflects the task related component of the human's behavior. Considering the observed human policy as a factored MDP [2], [3], the matched policy segregates task related (model fitting) behavior from that due to unrelated latent factors (noise). By identifying a set of task satisfying policies we keep the door open for moment to moment variations in intent (nonstationarity) while at the same time excluding extraneous behavior (noise), the two primary obstacles to modeling human behavior. The policy library provides a collection of 'types' to which a human trajectory is matched. The policy type provides a projection of an approximation of the human's task relevant behavior which can be used as a surrogate in predicting and responding to the human.

[1]School of Information Science, University of Pittsburgh, PA, USA. Email: hul52@pitt.edu, mlewis@sis.pitt.edu
[2]Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA. Email: tianwein, katia@cs.cmu.edu, siddhara, danahugh@andrew.cmu.edu

A. Motivating Example

A long standing observation in human factors is that people often perform even skilled tasks in very different ways. For example, [4] found that dispatchers controlling the British gas grid relied on widely varying heuristics and practices yet achieved roughly equivalent results. In order to assist humans who may perform the same task in different ways a robot must recognize how the human has chosen to perform the task and choose its own policy to maximize their joint performance. Consider a human and robot charged with loading a truck. The human could stand on the gate while the robot traveled back and forth from the warehouse to bring him boxes. Conversely, both robot and human could carry boxes from the warehouse to a staging area beside the truck and then quickly complete the task with the robot passing boxes from the cache up to the human in the truck or, they could switch half-way through and start packing boxes before getting more from the warehouse. If we could learn all the sets of human and robot policies capable of performing the joint task, we will have captured the task related components of human behavior. A robot observing the human carrying a box from the warehouse could match this behavior with a cache filling policy and, searching its own table, choose a complementary policy in which it also begins retrieving boxes rather than one that, for example, carries a box to the truck then waits for the

missing human. If the human were to take a break or leave for lunch these actions should not match any of the task-performing policies and hence decrease similarity across the set in a roughly uniform way, insulating the policy selection mechanism from error. Conversely, if the human were to switch from a cache-filling to a box-passing policy, similarity to cache-filling would decrease while that for box-passing would increase leading the robot to adopt a new complementary policy. It does all this without ever explicitly representing anything other than the similarity between a human trajectory and the policies in its library. The key to this approach lies in finding a model free measure of the similarity between noisy human trajectories and known policies.

B. Similarity Metric

We introduce the **cross-entropy metric** (CEM) as a policy similarity metric. Cross-entropy, often used as a loss function for classifiers, can measure the distance between two policies π_1, π_2:

$$\mathrm{CEM}(\pi_1, \pi_2) = E_{s,a\sim\pi_1}[\log \pi_2(a|s)] \tag{1}$$

where $\pi_1(\cdot|s), \pi_2(\cdot|s)$ are action distributions given state s.

If the policy π_2 and state-action samples from π_1 are obtained, we can then estimate the cross-entropy between two policies $\mathrm{CEM}(\pi_1, \pi_2)$ by Monte Carlo sampling, even without access to the target policy π_1. In human-agent teaming, human policy π_H, though unknown, can be compared using CEM as the similarity metric with each agent policy π_A in the library because the state-action pairs generated by the human policy can be obtained during interaction.

Given a sliding window of frames that record the observed behavior of the human policy π_H, we can estimate the CEM between a human policy π_H and any known agent policy π_A by the following formula:

$$\frac{1}{T}\sum_{t=1}^{T} \log \pi_A(a_t|s_t), \quad \text{where } (s_t, a_t)_{t=1}^{T} \sim \pi_H \tag{2}$$

where $(s_t, a_t)_{t=1}^{T}$ are the sequential state-action pairs from human policy play and T is the window size, a hyperparameter to be tuned.

II. Application

Team Space Fortress (TSF) is a cooperative computer game where two players control their spaceships to destroy a fortress [5]. A sample screen from the game is shown in Fig. 1. At the center of the screen lies a rotating fortress. The fortress and two spaceships can fire missiles towards each other when they are within the hexagonal area. The spaceship acting as *bait* enters the activation region first and tries to attract the fortress's attention. When the fortress attempts to shoot at the bait, its shield lifts making it vulnerable. The other player in the role of *shooter* can now shoot at the fortress and destroy it.

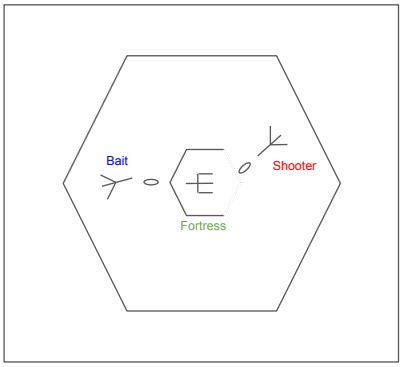


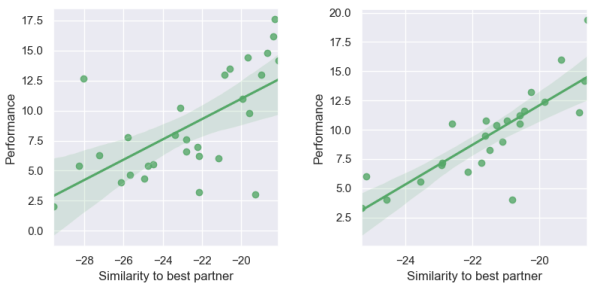
Figure 1: Sample TSF game screen

Reinforcement learning, rule based, and hybrid imitation/reinforcement learning polices were refined in self-play to produce policy libraries for the bait and shooter roles. These libraries were then played against each other to produce a paired-policy table containing the expected outcome of each pairing between bait and shooter policies. The highest scoring partner of a policy was designated that policy's complement.

We recruited 104 participants from Amazon Mechanical Turk for a human-agent experiment reported in [6]. Participants were randomly assigned a role of either shooter or bait and then teamed in randomized order with five artificial agents in the opposite role. Participants completed three 1-min game trials with each agent. The five variants were selected from our static agent library $\mathcal{L}$ balanced for performance in the self-play table and diversity by considering different training methods and reward functions.

To test the efficacy of our CEM similarity measure and its ability to predict human-agent team performance from play among surrogates, we compared the CEM dissimilarity between the complement of the policy matched to the human (best partner) and the policy with which the human played (tested partner).

Correlation analysis shows that "similarity to best partner" is positively correlated with team performance in both bait ($r = 0.636, p = .0002$) and shooter ($r = 0.834, p < .0001$) groups. This result in which complementary pairings of the human shooter accounted for 70% of the variance among teams shows the high payoff potentially available from our approach to matching. The result indicates that the complementary



(a) Human baits (b) Human shooters

Figure 2: Scatter plots of average team performance and "similarity to best partner" measurement of human players. Team performance is measured by average fortress kills per minute. Similarity is measured by CEM where negative values close to zero indicate a close distance between human policy and best partner policy. Solid lines and shaded area indicate linear regression estimations.

policy pairs we found in agent-agent self-play can be successfully extended to human-agent teams, and that our proposed architecture is able to accurately identify human policy types and predict team performance.

III. Conclusions and Future Work

In [6] we took the next step of testing an agent that adapted to its human partner. Our experiment employed two controls:

1) a condition in which the agent's policy was selected Randomly at each trial to control for improvements due to learning the task and interface
2) a condition in which the agent's policy was selected randomly on the first trial and remained Fixed throughout the remainder of the experiment to control for human adaptation to agent policy
3) Experimental condition: Adaptive Agent that dynamically chooses the best complementary policy

Figure 3 shows our results. In the initial trials as participants master the task, they have equivalent performance. During an interim period, the Adaptive Agent leads to significantly better team performance. In the final period, the Fixed policy group catches up with the Adaptive Agent while performance in the Random group remains back at the level it reached in the initial period.

The failure of the adaptive agent to maintain its advantage in later trials is troubling as our work was premised on improving human-agent teams through agent adaptation. Humans, however, adopted a variety of shooter policies choosing S11, the second worst shooter policy 17% of the time while never choosing either of the two best performing policies. By choosing a policy that performs poorly even when paired with its best complement, human players limit the potential performance of their teams so that humans adapting to some of the more favorable complements in the Fixed condition eventually caught up. As well as aiding adaptation, the policy library and cross-play table could serve as a roadmap for nudging human players with poor policy choices toward potentially better pairings. The CEM measure provides a way to measure distances among policy alternatives and serve as a basis for algorithms to estimate resistance and expected gains from shifting agent policies to induce changes in their human partner. The current research suggests that such work is needed and the problem of adaptation may best be addressed symmetrically.

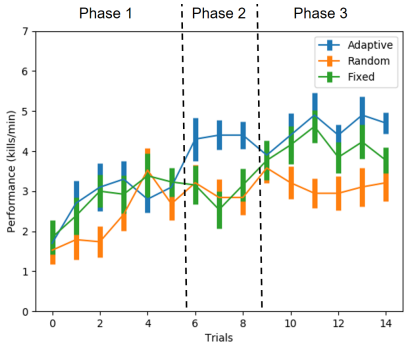


Figure 3: Human Shooters with Random, Fixed, and Adaptive Agent Baits

This research has additional limitations that could be improved by future work. First, the effectiveness of the proposed adaptive agent depends on the representativeness of the policy library. A larger or more precise coverage in the policy space of the team task could lead to more accurate estimation of human policy and better selection of complementary policies. In the present study agents were trained in plausible ways we thought likely to encompass actual human policies. In future work, we would like to enrich the static agent library using methods [7] designed to generate a diversity of policies providing assurance of coverage. Our method is dependent on clearly defined roles to exhaustively compare policy combinations to optimize the library. If boundaries between roles are porous (tasks can be performed by either agent) this process becomes much more difficult. In addition, required comparisons increase exponentially in the number of roles which along with diversity linked increases in number of policies could make optimizing computations expensive for larger problems. At execution, in compensation, comparisons are linear in the number of policies matching an actor's role.

Acknowledgment

This research has been supported by ARL DEVCOM DAC award W911NF-22-2-0001 and ARL award W911NF-19-2-0146

REFERENCES

[1] J. A. Bagnell, A. Y. Ng, and J. G. Schneider, "Solving uncertain markov decision processes," 2001.

[2] S. Nikolaidis, R. Ramakrishnan, K. Gu, and J. Shah, "Efficient model learning from joint-action demonstrations for human-robot collaborative tasks," in 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, 2015, pp. 189–196.

[3] V. V. Unhelkar, S. Li, and J. A. Shah, "Semi-supervised learning of decision-making models for human-robot collaboration." CoRL, pp. 192–203, 2019.

[4] I. Umbers, "A study of the control skills of gas grid control engineers," Ergonomics, vol. 22, no. 5, pp. 557–571, 1979.

[5] A. Agarwal, R. Hope, and K. Sycara, "Challenges of context and time in reinforcement learning: Introducing space fortress as a benchmark," arXiv preprint arXiv:1809.02206, 2018.

[6] H. Li, T. Ni, S. Agrawal, F. Jia, S. Raja, Y. Gui, D. Hughes, M. Lewis, and K. Sycara, "Individualized mutual adaptation in human-agent teams," IEEE Transactions on Human-Machine Systems, vol. 51, no. 6, pp. 706–714, 2021.

[7] J. Parker-Holder, A. Pacchiano, K. Choromanski, and S. Roberts, "Effective diversity in population based reinforcement learning," in Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20. International Joint Conferences on Artificial Intelligence Organization, 2020, p. 5923—5929.