Performance Boosting: Inverse Reinforcement Learning To Augment Rather Than Replace Humans

Mark Rucker

November 21, 2017

Abstract

Human-machine collaboration has the potential to solve a huge array of problems. Unfortunately, the path to this future is anything but clear. How much should machines replace human labor? How can human autonomy be respected while also protecting us from negative outcomes? How can human safety be ensured in these new collaborations? And even if these questions were resolved, how do we build these machines? This proposal lays out two simple experiments to explore how inverse reinforcement learning might offer a solution via a unique blending of human and computer, actions and goals.

1 Introduction

Human-machine collaboration has the potential to solve a huge array of problems. Unfortunately, the path to this future is anything but clear. How much should machines replace human labor? How can human autonomy be respected while protecting us from negative outcomes? How can human safety be ensured in these new collaborations? There is very little agreement, even among experts, on how to answer questions such as these.

Furthermore, even if these questions could be resolved, there still remains the incredible challenge of creating these collaborative machines. In the opinion of the author, this challenge – at an intuitive level – stems from four qualities required for collaboration:

1. **Goal Understanding**: What is it the human is trying to do?
2. **Action Selection**: Which action would be most helpful to the human?
3. **Action Coordination**: How should the selected action be communicated?
4. **Outcome Measurement**: What are the indicators of successful assistance?

The current state of the art in human-machine collaboration has, step by step, made progress through a combination of algorithm advances (c.f., [10, 18]) and careful problem framing (c.f., [4, 5]). This project proposes to make a similar – but still novel – contribution, with two experiments.

The first experiment will seek to show that a statistically significant relationship exists between an inverse reinforcement learning (IRL) reward function and human performance on a specific task. This connection has been suggested by the author's previous work [7] – which used IRL to replicate synthetic behavior – but was not directly measured in human subjects. The second experiment will seek to show that a statistically significant performance improvement can be measured when a simple computer assistant acts to improve performance.

Taken together these experiments will provide evidence to back up the reinforcement learning (RL) assumption: reward signals are the most concise and descriptive representations of behavior [10]. And from these experiments, future research can extend the ideas to more realistic human tasks (such as studying or editing a photo).

The remainder of the proposal proceeds as follows. Section 2 covers preliminary topics important to understanding the specifics of the proposal. Section 3 outlines the experimental designs, as well as risks to experimental validity and how to manage them. Section 4 details a timeline for the project. And section 5 concludes with contributions and future work.

2 Preliminaries

The proposed research project has been informed and motivated by existing research in IRL and human-machine collaboration. A basic review of these areas and their associated terminology is covered to help with understanding and to give the proposal context.

2.1 Preliminary to Inverse Reinforcement Learning

IRL is best understood as a sub-topic within RL. Given this relationship, a short review of RL is covered before proceeding to the IRL preliminary review.

RL is an area of machine learning motivated by the idea that intelligent behavior is best understood through the lens of reward seeking [2]. In RL, reward seeking behavior is modeled as a series of discrete sequential decisions in pursuit of a reward. The mathematical framework for these sequential decisions is often a Markov decision process (MDP).

RL, and thus IRL, grounds its approach in the pioneering research of behavioral psychologists such as Edward Thorndike and B.F. Skinner, who, in the early 20th century, provided convincing evidence that animal behavior can be understood simply as the pursuit of environmental rewards [17]. Recent advances in RL seem to further bolster reward driven approaches to behavior by leading to the creation of machines that can win against a Go world champion [3], perform complex helicopter acrobatics [1] and master Atari video games [9].

The RL approach can be thought of as a pipeline where inputs (reward functions) are fed into models (MDPs) to generate outputs (behavior policies). From this perspective IRL is the clear inverse of RL – behavior policies (inputs), are fed into models (MDPs) to generate reward functions (outputs) [16]. Many approaches to solving the IRL problem have been proposed: from linear programming [10], to quadratic programming [2, 15], to probability distributions [14, 21], to deep learning [18]. Using these algorithms a number of IRL applications have been explored: depression identification [6], pilot intent classification [20], cultural preference measurements [11] and human behavior prediction ([13], [8], [12], [19]).

2.2 Preliminary to Human-Machine Collaboration

The area of research concerning human-machine collaboration is large and not well defined. Even the name "human-machine collaboration" is primarily a creation of the author. Some of the research domains reviewed include: human-robot collaboration, persuasive technology, assistive agents, companion technology, adaptive interfaces and human-AI collaboration.

For the purpose of this review, only research which directly reasons about human agents is included. This means that tools – such as spam classifiers – are not reviewed even though they certainly collaborate with people in some sense. There are many commonalities within this corpus:

1. **Stochastic Models**: MDP, partially observable MDP, mixed observability MDP, decentralized-partially observable MDP

2. **Objective Inference**: Binary search, Bayesian prior with updates, IRL

3. **Psychological Theory**: Belief-Desire-Intention, Behavior cognitive theory, social cognitive theory, theory of planned behavior, rational choice theory, bounded memory

Perhaps, the primary take away is that there is relatively little agreement about the best way forward. Currently, much of the research is going on within two areas: robotics and persuasive technology. The reason is that both of these fields are testable and have more tangible applications than general assistance tools – such as planners and schedulers.

The experiments below borrow from what makes persuasive technology successful and applies it to human-machine collaboration. This reframes the collaboration problem from asking what can the computer do without interfering to what can the computer do to improve the human's performance.

3 Experimental Design

The proposal is to test two hypothesis:

1. IRL reward functions are predictive of task performance
2. IRL reward functions can lead to effective task assistance

3.1 Test For Hypothesis 1

To test hypothesis 1, a web interface will be created where human subjects will click on targets that appear on an otherwise blank page. Before beginning, subjects will be informed that their objective is to either click as close to the center of each target as possible (accuracy) or to click on as many targets as possible (speed). The accuracy objective will be measured by tracking the distance each click is from the center of a target with 0 being a perfect score for a click. The speed objective will be measured by counting the number of targets a subject is able to click on during the experiment.

Along with the objective measurements the features listed below (Table 1) will be tracked. An observation of the features will be collected every half second. These features will ultimately be used to calculate a reward function for each subject. It is important to note that the objective measurements (speed and accuracy) will not be included when calculating the reward function for a subject.

Reward Feature	**Objective**
Speed of Mouse	Speed
Mouse Location	Speed
Number of Clicks	Accuracy
Number of Targets	Accuracy

Table 1: Proposed Reward Features

At the conclusion of the experiment a short exit interview will be given to each participant with three questions: how well do you think you did, how enjoyable was the task and how difficult was the task.

Once all sessions are completed, and the reward functions have been calculated from the feature observations, the hypothesis will be tested. This test will be done in two steps. First, using hierarchical clustering with cosine distance and average linkage, clusters of reward functions will be created. From these clusters an ANOVA test will be performed to determine the significance of the clusters to performance. A significance of $p \leq .05$ will be used to reject the null.

A considered extension to this experiment is providing subjects certain customization options. Perhaps subjects could change the size or color of the targets. These kinds of personal preferences could provide greater insight into what each user finds rewarding. Particularly, if these controls could be changed as the experiment was ongoing, this would create a tension between taking the time to create a more productive environment and using that same time to simply be productive. Measuring the rewards that resolve this tension would be one more piece of information to explain performance.

3.2 Test For Hypothesis 2

To test hypothesis 2, the first experiment will be repeated with a new study population (it is important that the test participants for the second run are all new so better performance isn't caused by repetition). Everything will be the same, the same objectives will be given to study participants, the same objective measurements will be recorded and the same features will observed every half second.

What will be different with experiment 2 is an online IRL algorithm that calculates subjects' reward functions every half second and a digital assistant that works to help subjects perform better. Before going into detail about each of these difference, it should be said that comparing the performance of this second group (the group with the assistant) to the first group is the heart of this test.

The online IRL algorithm will likely require the development of a new algorithm. No existing algorithm that satisfied this requirement was found in a literature search. However, a number of existing algorithms do appear to be close. For example, simple modifications to the Bayesian inverse reinforcement learning algorithm by Ramachandran [14] may create an acceptable algorithm. Either way, this will need to be investigated as part of this project.

Apart from creating or modifying IRL algorithms, another option is to keep the state space small. Given the small number of features planned for the experiment (4), and appropriate discretization of their values (perhaps 10 levels each, though this may in practice turn out to be too few) the state space could be as small as $10^4 = 10,000$. This could allow the use of existing IRL algorithms (mentioned in the preliminary section) despite none being designed for online use.

For the assistant, it will act on its own without requiring any intervention from test participants. The assistant's goal will be to take actions so that subjects' online reward functions move closer to the highest performing reward function cluster found in test 1. A separate action for each feature has been selected (Table 2).

Action	**Objective**
Alter Mouse Speed	Speed
Alter Target Appearance Rate	Speed
Alter Mouse Clicks	Accuracy
Alter Target Color	Accuracy
Feedback on Performance	Speed/Accuracy

Table 2: Proposed Assistant Actions

Behaviorally, the assistant will follow a greedy policy. The action that corresponds to the reward feature most different from the high performing reward function will be taken. This naive assistance policy will be considered acceptable if participants' reward functions converge to the high performing reward function. If convergence is not seen, a more nuanced intervention policy may be needed. Convergence towards the desired reward is necessary to test the hypothesis that altering a participants reward functions can assist in task performance.

Finally, as with experiment 1, a short exit interview will be given to each participant with three questions: how well do you think you did, how enjoyable was the task and how difficult was the task.

Once all the results are gathered for experiment 2 Welch's t-test will be used to compare the mean performance (speed or accuracy measurements, respectively) of the entire sample population from experiment 2 to the mean performance of the entire sample from experiment 1. The null hypothesis for this experiment, intervening on reward functions does not effect performance, will be rejected if the test has $p \leq .05$.

3.3 Experimental Risks And Mitigation

One risk that will need to be managed are incorrect results due to execution errors in administering the two tests. In particular, experiment 2 needs to be executed with care given that it has many more pieces to it. This is, of course a risk with any experiment. For this project a detailed protocol will be created for both experiments ahead of time and carefully followed, along with quality assurance tests for all software.

A second risk that will need to be mitigated is the potential for type 2 errors. Changes in performance due to the assistant could be small, or variations between users could be large. Since each participant will be completing the experiment on their own computers in their own homes large variation is especially possible. Power calculations will need to be done ahead of time to minimize the chance that results are missed due to high variations or small differences. This process will include a review of existing literature on human task performance to get a sense of the expected effect size. With the effect size an appropriate sample size can be determined to reduce the probability that real effects are missed.

4 Plan Of Execution

The project will be organized and completed in three phases. In phase one all preparations will be made for the experiments, in phase two the experiments will be carried out and in phase three the data will be analyzed and final results will be written up. The expected timeline along with the components in each phase is below (Figure1, Table 3).

Effort has been made to front load work so unexpected problems can be addressed early.

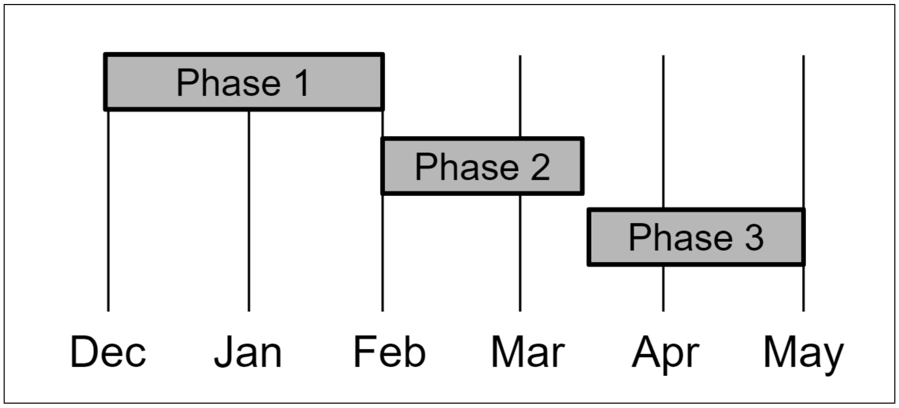

Figure 1: Projected timeline for project

Phase 1	IRB approval, task application creation, AWS setup for IRL, Amazon Mechanical Turk registration, development of online IRL algorithm
Phase 2	Test 1 execution and analysis, test 2 execution
Phase 3	Final analysis, paper and defense

Table 3: Tasks belonging to each project phase

5 Research Contributions

The project is anticipated to make three contributions to the research community, and the final results will be submitted to the journal of Computers in Human Behavior for publication to share these contributions:

1. Develop a new, online, IRL algorithm
2. Analyze an IRL reward function's relationship to performance
3. Analyze a digital assistant driven by an IRL reward function

For the UVA community specifically, this project will contribute to many existing research interests. Research regarding medical devices, industrial manufacturing and recommender systems could all potentially benefit from a better understanding of how humans and technology interact. IRL has the potential to provide a powerful new tool for all of these problems.

Perhaps the biggest weakness of the proposed project will be the remaining gap between experiment and general application. This gap comes from two limitations of the experiments:

the experimental task is unrealistically simple with clear actions for the assisting agent and the assisting agent has very little intelligence and no learning capacity. The first limitation means that, no matter the result (positive or negative), the result may not hold in more complex, real-world situations (e.g., a digital assistant to help individuals edit a photo). The second limitation means that, again, no matter the result, the results may not hold even for a more intelligent assisting agent. Building on this work and addressing these limitations is of interest to the author in future research.

References

[1] Abbeel, P., Coates, A., and Ng, A. Y. Autonomous helicopter aerobatics through apprenticeship learning. *The International Journal of Robotics Research 29*, 13 (2010), 1608–1639.

[2] Abbeel, P., and Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the 21st International Conference on Machine Learning* (2004), ACM, p. 1.

[3] Chen, J. X. The evolution of computing: Alphago. *Computing in Science & Engineering 18*, 4 (2016), 4–7.

[4] Fern, A., and Tadepalli, P. A computational decision theory for interactive assistants. In *Advances in Neural Information Processing Systems* (2010), pp. 577–585.

[5] Fitzpatrick, K. K., Darcy, A., and Vierhile, M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial. *JMIR Mental Health 4*, 2 (2017), e19.

[6] Huang, H., Harlé, K., Movellan, J., and Paulus, M. Using optimal control to disambiguate the effect of depression on sensorimotor, motivational and goal-setting functions. *PLOS ONE 11*, 12 (2016), e0167960.

[7] Lee, Rucker, S. B. G. K. Agent-based model construction using inverse reinforcement learning. In *Proceedings of the 49th conference on Winter simulation* (2017), winter simulation conference.

[8] Liu, S., Araujo, M., Brunskill, E., Rossetti, R., Barros, J., and Krishnan, R. Understanding sequential decisions via inverse reinforcement learning. In *Proceedings of the 2013 IEEE 14th International Conference on Mobile Data Management* (2013), vol. 1, IEEE, pp. 177–186.

[9] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).

[10] Ng, A. Y., and Russell, S. J. Algorithms for inverse reinforcement learning. In *Proceedings of the 17th International Conference on Machine Learning* (2000), pp. 663–670.

[11] Nouri, E., Georgila, K., and Traum, D. R. A cultural decision-making model for negotiation based on inverse reinforcement learning. In *Proceedings of the Cognitive Science Society* (2012).

[12] Osogami, T., and Raymond, R. Map matching with inverse reinforcement learning. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence* (2013), AAAI, pp. 2547–2553.

[13] Qiao, Q., and Beling, P. A. Recognition of agents based on observation of their sequential behavior. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (2013), pp. 33–48.

[14] Ramachandran, D., and Amir, E. Bayesian inverse reinforcement learning. *Urbana 51*, 61801 (2007), 1–4.

[15] Ratliff, N. D., Bagnell, J. A., and Zinkevich, M. A. Maximum margin planning. In *Proceedings of the 23rd international conference on Machine learning* (2006), ACM, pp. 729–736.

[16] Russell, S. Learning agents for uncertain environments. In *Proceedings of the 11th annual Conference on Computational Learning Theory* (1998), ACM, pp. 101–103.

[17] Sutton, R. S., and Barto, A. G. *Reinforcement Learning: An Introduction.* Cambridge Univ Press, 2011.

[18] Wulfmeier, M., Ondruska, P., and Posner, I. Maximum entropy deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888* (2015).

[19] Yang, S. Y., Qiao, Q., Beling, P. A., Scherer, W. T., and Kirilenko, A. A. Gaussian process-based algorithmic trading strategy identification. *Quantitative Finance 15*, 10 (2015), 1683–1703.

[20] Yokoyama, N. Intent inference of aircraft via inverse optimal control including second-order optimality condition. In *AIAA Guidance, Navigation, and Control Conference* (2017), p. 1254.

[21] Ziebart, B. D., Maas, A. L., Bagnell, J. A., and Dey, A. K. Maximum entropy inverse reinforcement learning. In *Proceedings of The Twenty-third AAAI Conference on Artificial Intelligence* (2008), pp. 1433–1438.