Learning from Single and Multi-human intelligence via Cognitive and Reinforcement Learning

Aadhar Gupta S21007@STUDENTS.IITMANDI.AC.IN
Mahavir Dabas MAHAVIRDABAS18@GMAIL.COM
Shashank Uttrani SHASHANKUTTRANI@GMAIL.COM
Varun Dutt VARUN@IITMANDI.AC.IN
Applied Cognitive Science Lab, Indian Institute of Technology Mandi, Kamand, Himachal Pradesh, India – 175005

Abstract

Prior research has explored reinforcement learning (RL) and cognitive algorithms to learn from scratch or via human demonstrations, but research currently lacks an investigation of collective demonstrations of multiple humans to drive artificial agents. This paper presents and evaluates multi-human demonstration driven Instance-based Learning (IBL) and Deep Q-Learning Network (DQN) models for playing the cart-pole game. Both IBL and DQN models were primed with single or multi-human training data and then made to perform in test scenario in cart-pole. Results revealed that IBL and the DQN model performance improved by around 177% and 30%, respectively, when driven by multi-human demonstrations compared to single-human demonstrations. Results also revealed that IBL far outperformed the DQN model in game-play, by learning generalized useful and goal-oriented actions from behavior of multiple humans.

1. Introduction

Reinforcement Learning (RL) is a paradigm of machine learning where the agent learns with the help of the feedback signal (rewards) received corresponding to various environmental interactions (Sutton, 1992). Contrary to the supervised learning, RL was conceived to be used when the target to achieve was unknown. The model explores the various possible alternatives and hence discovers the most optimal one, also known as exploration and exploitation of knowledge (Sutton, 1992). Q-learning (Watkins & Dayan, 1992) is a model-free RL algorithm (Sutton, 1992), that, instead of relying on pre-determined transition probabilities and reward functions governing the decisions, uses a trial-and-error approach to interact with the environment to approximate the optimal decisions. The algorithm aims to determine the State-Action values (Q-Value), i.e., the expected discounted sum of cumulative future rewards corresponding to various actions taken at various states (Gupta et al., 2021) and stores them in a table called the Q-table (Watkins & Dayan, 1992).

The marriage of neural networks with Reinforcement Learning gave rise to Deep Reinforcement Learning (DRL) capable of generalizing over vast state-action spaces. DRL also paved the way for application of RL to complex problems, previously regarded intractable (Arulkumaran et al., 2017). The recent success of RL in beating human performance in tasks like playing Atari games

(Mnih et al., 2015) and defeating the world Go champion (Borowiec, 2016) have demonstrated the capability and robustness of RL algorithms. DQN (Arulkumaran et al., 2017) is an upgraded version of Q-learning, where instead of maintaining a table to store all the possible Q-values, a neural network is used to learn the various Q-values. This helps deal with the problems with state-action space too large or impossible (boundless state-action space) to be contained within a Q-table.

The DRL techniques, however, suffer from some significant shortcomings such as reward shaping, sample inefficiency, and local optima due to high complexity (dsh, 2019). Moreover, learning for sparse reward systems has its own set of challenges (dsh, 2019). Although prior research has attempted to address these issues, (Pathak et al., 2017), another approach to teach optimal behavior to the agent is through human demonstrations. This notion lays the foundation of the field of Imitation Learning (IL) (Schaal, 1999) and some aspects of cognitive modeling (Katakwar et al., 2022b,a; Kotseruba et al., 2016; Kotseruba & Tsotsos, 2020; Ye et al., 2018). IL is a branch of AI where the agent tries to mimic human behavior by copying the human decision at the closest state (Schaal, 1999). Cognitive modeling is a branch of Artificial Intelligence (AI) that aims at creating techniques as robust, insightful, and adaptive as human intelligence (Kotseruba & Tsotsos, 2020).

Prior research has contributed to more than a hundred cognitive architectures (Chong et al., 2007; Kotseruba & Tsotsos, 2020) mainly falling under the following categories: symbolic representation and production rule-based inference (Laird et al., 1987), psychology-based models to mimic human cognition (Anderson et al., 2004; Langley & Choi, 2006), incorporating beliefs, desires, and intentions (Bratman et al., 1988); and combining neural networks with cognitive psychology (Sun & Peterson, 1996). Adaptive Control of Thought-Rational (ACT-R) (Anderson et al., 2004) is a psychologically motivated cognitive model that combines AI, cognitive psychology, and some components of neurobiology. Many researchers have extended upon the principles of ACT-R yielding architectures avoiding the high complexity yet retaining the efficiency, such as IBL (Gonzalez et al., 2003; Gonzalez & Dutt, 2010, 2011). The IBL model is a cognitive framework very well suited to address decision-making problems. In IBL, the information of past experiences is stored in the form of situation-decision-utility (SDU) triplets called instances (Gonzalez et al., 2003) and uses them to predict the optimal decision for the current scenario (Gonzalez & Dutt, 2010). The expected utility, i.e., the expected reward corresponding to each decision alternative, called the blended value (BV), is computed, and the decision corresponding to the maximum BV is chosen. Various cognitive models have been studied over a wide range of autonomous tasks including robotics, computer vision as well as playing games of Freeciv, Atari Frogger II, Infinite Mario, browser games, and Backgammon (Kotseruba & Tsotsos, 2020).

Games provide portable, cheaper and more customizable alternatives to real-world scenarios for testing of AI algorithms. As a result, a large volume of AI research revolves around games (Borowiec, 2016; Kotseruba & Tsotsos, 2020; Mnih et al., 2015), (Singal et al., 2017). The cartpole game has been used as a benchmark problem to test a number of AI algorithms for control (Kumar, 2020; Nagendra et al., 2017). The game environment consists of a pole attached to the cart like an inverted pendulum, tending to tumble, and the player needs to balance the pole by moving the cart.

Previous works have investigated various RL(Singal et al., 2017) and cognitive approaches(Kotseruba & Tsotsos, 2020) for playing games. Besides learning via trial and error, another widely used approach for developing AI agents is learning by watching/demonstration (Abbeel & Ng, 2004; Ho

& Ermon, 2016; Miyashita et al., 2017; Torabi et al., 2018). However demonstrations of a human naturally embrace personal biases and individual-specific noise (like some temporary or permanent distraction influencing the cognition). These irrelevant components of data might unnecessarily increase the complexity of data: confusing the model and preventing it from capturing the desirable decision-making patterns. Moreover, each individual possesses a particular style of behavior leading to redundancy in behavior patterns that may limit how much the model could gain from a single individual's behavior demonstrations. On the contrary, the combined demonstrations of multiple humans could suppress personal biases and individual-specific noise. However, literature lacks investigation of multi-human demonstrations for driving AI agents. For example, very little is known on how data from multiple human sources could be used in synchronization to lead to a better test performance of cognitive models.

To address this literature gap, in this study, we attempted one of the numerous possible ways of combining multi-human demonstrations. The experiment began with the development of a cart-pole game environment. Next, ten human participants were recruited to play the game and generate human game-play demonstrations. Subsequently, IBL and DQN models were developed using the human demonstrations obtained. First, the models driven by single human player demonstrations were developed, followed by the models based on multi-human player demonstrations. The models were evaluated on the basis of the score earned on playing the cart-pole game. The game-play results of models were analyzed and interpretations were drawn.

The remaining paper is structured as follows: First the methodology is described, covering the experiment design, data collection procedure, model building, and the model evaluation procedure. Next, the results are presented, interpreted, and the significance of the findings are discussed. Eventually under the conclusion section, the findings are summarized, followed by a mention of the limitations of the study, the future scope of work based on this study, and the implications of this work for the AI community.

2. Methodology

2.1 Experiment Design

A cart-pole environment was developed using the PYGame python library (Pygame, 2000). The task was to balance the pole on the cart for as long as possible. There were two terminating conditions for the game: the angle of the pole with the vertical axis exceeding a threshold of 30 degrees; or, the cart falling off the platform. A reward of 0.1 and -5 was given for non-terminating and terminating actions, respectively. The game-play was divided into two phases: the training phase, with ten trials per player, and the testing phase, with a single trial per player. The situation of the cart-pole was defined by four values: cart position, cart velocity, pole angle to the vertical axis, and pole angular velocity. There were two possible actions for a participant to be taken in the game: left or right.

2.2 Data Collection

To collect the human game-play demonstration for cart-pole, 42 human volunteers were recruited from a reputed University in India. The data of ten highest scoring players was shortlisted for this study. The male and the female participation in the shortlisted ten players was 60% and 40%, with a count of 6 and 4 players, respectively. The age of the shortlisted players lay in the interval of 24 to 28 years (mean age = 25.5 years; standard deviation = 1.43 years). The participants were graduates spread across the various disciplines of technology and engineering: electronics, computer science, electrical and mechanical. None of the participants had any prior experience of the cart-pole game. The experiment began by instructing the participants about the game's rules. Each individual was given eleven trials on the game, ten for training and last one for testing the human. The score of the test trial was used as a metric to select the best ten players for this study. The actions taken by the players, along with the corresponding situation vector and the reward received, were recorded.

2.3 Model Building

2.3.1 IBL

Conceptual Details: In IBL, the computation of Blended Value is a multi-step process. Firstly, the instances relevant to the current situation are shortlisted against a similarity threshold. Then the Activation of each experience, analogous to weightage, is computed using the frequency and recency of its occurrence, its similarity with the current situation, and a random noise component, given as:

$$A_{i,t} = \sigma \ln(\frac{\gamma_{i,t}}{1-\gamma_{i,t}}) + \ln(\sum_{t_p=1}^{t-1} (t-t_p)^{-d}) + \mu(S) \tag{1}$$

where d represents the memory decay parameter, σ represents cognitive noise parameter to take into account the agent-to-agent variability in activations, γ is a random draw from a uniform probability distribution, t_p represents the index of time steps of the previous occurrence of instance i, S represents the similarity measure between the situation of instance i and the current test situation, and μ is the scaling factor, that always takes a positive value.

The activation of the shortlisted instances is used to compute the probability of retrieval of the instances, which is equivalent to the relevance of each experience for the current situation, computed as:

$$P_{i,t} = \frac{e^{A_{i,t}/\tau}}{\sum_j e^{A_{j,t}/\tau}} \tag{2}$$

where τ represents the random noise and $A_{i,t}$ represents the activation of the instance i.

Eventually, the Blended Value for an action j is computed using the relevance and the Utility of the shortlisted instances, given as:

$$V_j = \sum_{i=1}^{n} p_i x_i \tag{3}$$

where x_i represents the utility and p_i represents the probability of retrieval (PR) of the instance i.

Implementation: For the computation of activation value, the base activation did not play any role as all the memory instances were timestamped simultaneously (t=0). The activation value was completely dependent on the similarity value. Cosine similarity was used as the similarity metric.

An IBL model can learn optimal behavior in two ways: model gaining experience by interactions with the environment or model getting initialized with human experience (model gaining experience via human interactions with the environment). In this study, the IBL models were pre-populated with human training game-play demonstration, to follow human tactics in order to exhibit more human-like behavior. Two types of models were developed based on the number of human sources contributing the model's initial data: Single-human IBL model and Multi-human IBL model.

1. *Single-human IBL model:* The SDU tuples of the training game-play session of a single human player were used to initialize the IBL agents. A separate, dedicated IBL model was created corresponding to the data of each human participant. Hence there were a total of ten distinct single-human IBL models.

2. *Multi-human IBL model:* For the multi-human IBL model the data used to initialize the memory was collected from ten human players. The experiences of these human players were combined by sampling one-tenth of the behavior SDU instances from the data of each player. The sample size of one-tenth ensured that the multi-human data size did not exceed the average data size of the involved ten players, allowing the analysis of the impact of diversity in sources of data independent of the data size.

Hyper-parameters: For both the kinds of IBL models, the hyper-parameters used were the cognitive noise and similarity. Cognitive noise (CN) is added to capture the variability in decisions from one agent to another. For this study, the CN was varied over the range 0 to 1, and was eventually set to 0.25. Similarity term is used to adjust the weight of a memory instance depending on its relation/closeness with the current situation. Cosine similarity was used as the similarity metric in this study.

2.3.2 DQN

Conceptual details: In Q-Learning, the Q-values are updated using the Bellman equation, given as:

$$Q(s_t, a_t) = (1 - \alpha)Q(s_t, a_t) + \alpha(r_t + \gamma \max_a(Q(s_{t+1}, a))) \tag{4}$$

where $Q(s_t, a_t)$ represents the Q-Value for state s_t and action a_t, s_t and s_{t+1} stand for the current and the next state respectively, r_t represents the reward on the transition from the state s_t to the state s_{t+1} on taking action a_t, α represents the learning rate that controls the amount of difference between the new and old versions of the Q-Values, considered for updating the Q-values and γ represents the discount factor.

Implementation: The experience replay buffer (Watkins & Dayan, 1992) of DQN was initialized with the quadruplets of 'state, decision, feedback, and next state' from the human training session data. This enabled the model to learn from human behavior. The models were trained for 5+ epochs but no improvement was seen in the model prediction performance after 5 epochs. The models predicted two Q values corresponding to the actions, left and right. Like the IBL model, a distinct model instance was trained on each human player's data. A total of 11 DQN models were developed corresponding to the single and multi-human demonstration data:

1. *Single-human DQN model:* The experience replay buffer of this class of DQN models were directly initialised with the 'state, decision, reward, and next state' tuples from the recorded human game-play of the ten training sessions. The entire data of training trials was fed to the model. A total of ten such models were developed, one corresponding to the game-play data of each human player.

2. *Multi-human DQN model:* For multi-human model, the experience replay buffer was initialised with the 'state, decision, reward, and next state' tuples sampled from the game-play demonstration of each of the ten human human players. Again the proportion of data sampled from each player's demonstration was one-tenth. Similar to IBL, just one DQN model was developed for multi-human demonstration.

Network Architecture: The neural network architecture (16-32-2) comprised two fully connected hidden layers with 16 and 32 units and the output layer with two units corresponding to the two actions. Rectified linear activation followed by Dropout with a rate of 0.1 was used for both hidden layers. This architecture was arrived at by varying the number of nodes in the two hidden layers in the range: 1, 16, 32, and 64.

Hyper-parameters: Three hyper-parameters were associated with the model, the learning rate, the batch size and the discount factor (gamma). The learning rate controls the rate of updation of the Q-values, and was set to 0.001. The batch size was set to 32. This batch size was arrived at by varying the batch size in the range: 8, 16 and 32. The discount factor, which determines the fraction of future rewards considered, was set to 0.95. This value for discount factor was arrived at by varying the discount factor in the range: 0.8, 0.9, 0.95, 1.0.

2.4 Model Evaluation

The IBL and DQN models were applied on the cart-pole agent and allowed to control the motion of the cart. In total, 22 models were evaluated; 11 models each for IBL and DQN including ten single-human models and one multi-human model. As the demonstration fed to the models was human players' attempt to maximize the game score by preventing the pole falling over for as long as possible, hence, the true test of the model learning is via the cart-pole game score. Each action taken in the game before the termination, awarded a score of 0.001. Hence, a score of 1 stated that the player managed to balance the pole for 1000 steps, before the termination condition was reached. The game score hence revealed the ability of the model to understand the context of the game, and, the model's versatility, that is, robustness in handling the various situations in the game.

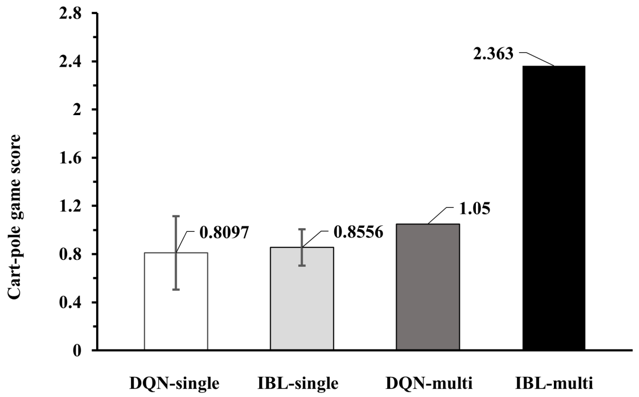

Figure 1. Comparison between the game scores of all the models developed: single-human DQN model (DQN-single), multi-human DQN model (DQN-multi), single-human IBL model (IBL-single) and multi-human IBL model (IBL-multi)

3. Results

This section provides the findings of experiments carried out to evaluate and compare the performance of various models developed in this study.

Figure 1 presents the performance of all the categories of models developed in this research, covering the single-human DQN model, single-human IBL model, multi-human DQN model and the multi-human IBL model. As shown in the figure, the y-axis represents the score earned in the cart-pole game during the test, and the x-axis depicts the type of the model. The average single-human DQN and IBL model performance was obtained using the mean score earned in cart-pole game-play, across the ten single-human model instances, each developed using the demonstrations of a distinct human player. In the figure, the average single-human DQN model performance is depicted through the white-coloured bar (DQN-single), and the average single-human IBL model performance is depicted through the light-gray-coloured bar (IBL-single). The error bars represent the 95% confidence interval, with a value of 0.3045 for DQN and 0.151 for IBL. As indicated in the figure, both the approaches produced very similar average game scores, with IBL outperforming DQN with a very small difference of 0.046 in the game score.

The dark-gray-coloured bar (DQN-multi) in the figure corresponds to the game score during test of the DQN model developed using combined demonstration of multiple human-players (multi-intelligence DQN model). It can be observed from this figure that the multi-intelligence DQN model scored relatively better in the cart-pole game than the average score of single-human DQN models. The multi-human DQN model reported a gain of 30 percent in the performance over the single-human DQN model, achieving a score of 1.05, while the average score for single-human DQN model being 0.809.

The black-coloured bar (IBL-multi) in the figure corresponds to the performance of the multi-intelligence IBL model. As one can see, a humongous improvement in performance was reported by the IBL model driven by the combined demonstration of multiple human-players (multi-intelligence IBL model), compared to the average score of the single-human IBL model. The multi-intelligence IBL model achieved a score of 2.36, reporting an increase of 177 percent over the average score of 0.856 for the single-human IBL models, almost tripling the value. The plot gives a clear indication of the improved reliability and robustness in the decision making for the multi-intelligence model, yielding lesser faults and a better control across a wide variety of states of the cart-pole game.

As it can be observed, the diverse multi-human demonstration yielded superior results compared to single-human demonstration, for both the approaches, DQN and IBL. Interestingly, as indicated in the figure, the result of IBL is significantly better than that of DQN with respect to learning from multi-human demonstration. It was found that the game score obtained by the multi-intelligence IBL techniques, 2.36 is 125% higher than that obtained by its DQN counterpart. It is therefore clear that IBL cognitive approach outperforms DQN reinforcement learning approach in harnessing multi-human demonstrations to extract the optimal behavior strategies.

4. Conclusion and Discussion

This investigation primarily focused on evaluating the effect on the performance of AI agents by the usage of behavior-demonstration-data sampled from multiple human sources. The study began with the recruitment of human players to play the cart-pole game. The players' behavior in the game was recorded in the form of instance-tuples consisting of the situation vector (current state of the game), corresponding human decision and the utility (reward obtained). The players were given 11 attempts in the cart-pole game. The first ten constituted the training session where the players learn the skills by trial and error, gradually gaining experience and improving. The final attempt constituted the test phase, where the player was expected to have gained expertise in the game and give the best possible performance. In this study the human game-play demonstrations from the test phase were used to develop the model, assuming it to possess the best possible version of the humans' game-play behavior. The demonstrations of each human player were used to pre-populate the IBL agents' memory and DQN agent's experience buffer. A distinct copy of the models was set up corresponding to each human player's demonstrations. The trained DQN agent and the initialized IBL agent were evaluated over their performance in cart-pole game. Both the models, IBL and DQN were found to yield similar performance, with IBL slightly edging over DQN. Subsequently, we experimented with a mechanism to enhance the human behavior modeling by increasing the diversity in human demonstrations, while maintaining equivalent demonstration data size. In this paper, we demonstrated how to enrich the human-behavior demonstration for AI agents by random sampling from the demonstrations of multiple human individuals. Results revealed that the demonstration data obtained by sampling from multiple human sources leads to a much better learning and improvement in performance, compared to a single-human's demonstrations. Moreover, the IBL model far outperformed the DQN model when driven by multi-human demonstration data. This shows the capability of IBL to capture the beneficial behavior patterns by generalizing over the behavior information from multiple humans, is superior to that of the DQN model driven by the same, multi-human behavior demonstrations.

This paper analyzed two AI models on human demonstrations for playing the cart-pole game, DQN representing reinforcement learning, and IBL representing cognitive modeling. Results revealed that when developed over a human individual's behavior demonstration, both the models produced similar performance in the game. It is clear that the model performance varies significantly with the demonstration data used, due to the difference in the thought process, strategies, response time and other factors influencing human behavior. However, the much larger inter-model variation corresponding to the different human sources, in DQN, highlights its sensitivity to changes in human behavior patterns. On the other hand, IBL appears to be a little more robust with respect to inter-human behavior variations. A likely reason could be the different approaches in which these two models deal with the human demonstration; the IBL cognitive model revolves around the immediate benefit whereas the DQN reinforcement learning technique focuses on maximizing the long term benefit.

Results also revealed a positive influence of multi-human demonstrations on the learning and performance outcome of AI agents. For both the approaches, IBL and DQN, the multi-intelligence model (multi-human demonstration-based) succeeded in outperforming the standard (single-human demonstration-based) counterparts. Interestingly, the IBL multi-intelligence model performed exceedingly better than the standard IBL model. When powered by multi-human demonstrations, the models seem to choose the more sensible option that is the one which has been taken the most often or has contributed the greatest to overall score. Safer and wiser behavior is likely to be encouraged when the model gets a glimpse of behavior of more than a single individual. Perhaps, these results indicate the presence of bias and individual-specific noise in a human individual's decision which may cloud the model's understanding of optimal behavior and hamper proper judgment. This result agrees with the notion that consulting a counsel of people yields better and more informed decisions than consulting a single person.

Though there was a clear improvement in the performance of both, IBL and DQN for multi-human demonstrations, there was a vast difference in the way these two approaches harnessed the information about the general human behavior lying in the combined demonstrations. IBL preferred choosing the better options more commonly than the DQN model. It might simply point to the DQN model's inability to prevent the immediate action or the sequence of action leading to termination. This appears to be very well handled by the underlying architecture of the IBL cognitive model. These results also indicate that the IBL cognitive model is better suited to filter out the biases and individual-specific noises and direct the focus towards common, goal-directed behavior. with presence bias and individual-specific noise in a human individual's decision which may cloud the model's understanding of optimal behavior and hamper proper judgment. Therefore, we can say that in general, AI models get benefitted by diversity in sources of the demonstration data, and cognitive models such as IBL are very well suited to learn optimal behavior by generalizing over such varying behavior patterns.

One of the possible explanations for remarkably good performance of IBL compared to DQN could be the inbuilt greedy approach leading to the choice of the most suitable action at any instant. In scenarios like cart-pole, the human player's decisions are primarily instantaneous. There is no significant role of sequential information, as the player chooses to go left or right based on the instantaneous situation of the pole. Thus, DQN'S long-sighted approach doesn't seem to work as effective as the IBL short-sighted approach for multi-human demonstrations. This is also consistent with the slightly better performance of IBL over DQN for single-human demonstration.

This experiment was subject to certain limitations due to the sampling procedure employed for combining the information from distinct human player's demonstration data. The multi-human demonstrations used in this study were prepared by sampling one-nth of each human player's demonstration data (n being the total number of humans). However, this study does not explore the optimal size of data to be sampled from each human's demonstrations, or if there exists a subset of human players within the total players, that leads to the optimal combined multi-human demonstration data. Additionally, although random sampling proved to be successful in improving the performance of the AI agents used, it is not a very logical and reliable way to extract information from each human's behavior demonstration. As sampling doesn't guarantee the extraction of best possible information, further investigation may be carried out on more intuitive and better informed methods to select experiences from a human demonstration, to create the multi-human demonstration.

This research opens up a wide scope for further investigations. Our work may be further extended by involving the state of the art techniques from cognitive modeling such as Soar and Clarion, and from the field of imitation learning such as GAIL and BC, to reveal the best performing technique based on human demonstrations and multi-intelligence concepts presented in this paper. Various ensemble models merging the ideas of multiple cognitive models or cognitive models with RL/IL models may be explored for multi-intelligence. As sampling doesn't guarantee the extraction of best possible information, further investigation may be carried out on more intuitive and better informed methods to select behavior instances from a human demonstration, to create the multi-human demonstration. Some possible directions would be assignment of weights to the behavior instances and subsequent shortlisting, or extraction of the behavior patterns most commonly occurring across human individuals. Moreover, for the development of IBL models in this research, only the hyper parameters of cognitive noise and similarity were considered. Due to initialisation of agent memory with the entire human behavior demonstration, each action got time stamped with zero, nullifying the usage of any sequential information. Inclusion of the base activation, built on the concepts of recency and frequency, may lead to further improvement in IBL models' performance. We plan to continue exploring in these directions in our ongoing research on multi-intelligence for artificial agents and IBL cognitive modeling for replicating human-like behavior.

References

Abbeel, P., & Ng, A. Y. (2004). Apprenticeship learning via inverse reinforcement learning. *Proceedings of the twenty-first international conference on Machine learning* (p. 1).

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological review*, *111*, 1036.

Arulkumaran, K., Deisenroth, M. P., Brundage, M., & Bharath, A. A. (2017). A brief survey of deep reinforcement learning. *arXiv preprint arXiv:1708.05866*.

Borowiec, S. (2016). Alphago seals 4-1 victory over go grandmaster lee sedol. *The Guardian*, *15*, 6.

Bratman, M. E., Israel, D. J., & Pollack, M. E. (1988). Plans and resource-bounded practical reasoning. *Computational intelligence*, *4*, 349–355.

Chong, H.-Q., Tan, A.-H., & Ng, G.-W. (2007). Integrated cognitive architectures: a survey. *Artificial Intelligence Review*, *28*, 103–130.

dsh (2019). Top 4 reasons why reinforcement learning sucks (ep. 83). `https://datascienceathome.com/what-is-wrong-with-reinforcement-learning/`. (Accessed on 06/23/2022).

Gonzalez, C., & Dutt, V. (2010). Instance-based learning models of training. *Proceedings of the human factors and ergonomics society annual meeting* (pp. 2319–2323). SAGE Publications Sage CA: Los Angeles, CA.

Gonzalez, C., & Dutt, V. (2011). Instance-based learning: integrating sampling and repeated decisions from experience. *Psychological review*, *118*, 523.

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science*, *27*, 591–635.

Gupta, A., Roy, P. P., & Dutt, V. (2021). Evaluation of instance-based learning and q-learning algorithms in dynamic environments. *IEEE Access*, *9*, 138775–138790.

Ho, J., & Ermon, S. (2016). Generative adversarial imitation learning. *Advances in neural information processing systems*, *29*.

Katakwar, H., Aggarwal, P., Maqbool, Z., & Dutt, V. (2022a). Influence of probing action costs on adversarial decision-making in a deception game. In *Ict analysis and applications*, 649–658. Springer.

Katakwar, H., Uttrani, S., Aggarwal, P., & Dutt, V. (2022b). Modeling the effects of network size in a deception game involving honeypots. In *Cybersecurity and cognitive science*, 339–355. Elsevier.

Kotseruba, I., Gonzalez, O. J. A., & Tsotsos, J. K. (2016). A review of 40 years of cognitive architecture research: Focus on perception, attention, learning and applications. *arXiv preprint arXiv:1610.08602*, (pp. 1–74).

Kotseruba, I., & Tsotsos, J. K. (2020). 40 years of cognitive architectures: core cognitive abilities and practical applications. *Artificial Intelligence Review*, *53*, 17–94.

Kumar, S. (2020). Balancing a cartpole system with reinforcement learning–a tutorial. *arXiv preprint arXiv:2006.04938*.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial intelligence*, *33*, 1–64.

Langley, P., & Choi, D. (2006). A unified cognitive architecture for physical agents. *Proceedings of the National Conference on Artificial Intelligence* (p. 1469). Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Miyashita, S., Lian, X., Zeng, X., Matsubara, T., & Uehara, K. (2017). Developing game ai agent behaving like human by mixing reinforcement learning and supervised learning. *2017*

18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD) (pp. 489–494). IEEE.

Mnih, V., et al. (2015). Human-level control through deep reinforcement learning. *nature*, *518*, 529–533.

Nagendra, S., Podila, N., Ugarakhod, R., & George, K. (2017). Comparison of reinforcement learning algorithms applied to the cart-pole problem. *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 26–32). IEEE.

Pathak, D., Agrawal, P., Efros, A. A., & Darrell, T. (2017). Curiosity-driven exploration by self-supervised prediction. *International conference on machine learning* (pp. 2778–2787). PMLR.

Pygame (2000). About - pygame wiki. `https://www.pygame.org/wiki/about`. (Accessed on 10/6/2022).

Schaal, S. (1999). Is imitation learning the route to humanoid robots? *Trends in cognitive sciences*, *3*, 233–242.

Singal, H., Aggarwal, P., & Dutt, V. (2017). Modeling decisions in games using reinforcement learning. *2017 International Conference on Machine Learning and Data Science (MLDS)* (pp. 98–105). IEEE.

Sun, R., & Peterson, T. (1996). Learning in reactive sequential decision tasks: The clarion model. *Proceedings of International Conference on Neural Networks (ICNN'96)* (pp. 1073–1078). IEEE.

Sutton, R. S. (1992). Introduction: The challenge of reinforcement learning. In *Reinforcement learning*, 1–3. Springer.

Torabi, F., Warnell, G., & Stone, P. (2018). Behavioral cloning from observation. *arXiv preprint arXiv:1805.01954*.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, *8*, 279–292.

Ye, P., Wang, T., & Wang, F.-Y. (2018). A survey of cognitive architectures in the past 20 years. *IEEE transactions on cybernetics*, *48*, 3280–3290.