Adversarial Inverse Learning of Defense Policies Conditioned on Human Factor Models

Amirhossein Ravari[1], Guangyu Jiang[2], Zuyuan Zhang[2], Mahdi Imani[1], Robert H. Thomson[3], Aryn A. Pyke[3], Nathaniel D. Bastian[3], Tian Lan[2]

[1]Northeastern University, Boston, MA
[2]George Washington University, Washington, DC
[3]United States Military Academy, West Point, NY
Emails: {ravari.a, m.imani}@northeastern.edu, {guangyu.jiang, zuyuan.zhang, tlan}@gwu.edu, {robert.thomson, aryn.pyke, nathaniel.bastian}@westpoint.edu

***Abstract*—The rapid growth of wireless networks has led to a significant increase in cybersecurity threats. Learning defense policies from offline data and expert demonstrations can eliminate the need to interact with a network/simulator during training, producing intelligent agents for automated/assistive defense. However, existing works on imitation and inverse learning are oblivious to various human factors that often affect decision-making and workflow in network defense. Learning defense policies that are conditioned on human factors can significantly improve the effectiveness of learned policies/models in network defense, better reflecting the varied behavior of human operators. To this end, we propose an adversarial inverse reinforcement learning algorithm incorporating models of human factors to learn user-specific policies/models in cyber defense from demonstrations and trajectories. The proposed algorithm captures varied operator behaviors through the modeling of three human factors: fatigue, expertise, and risk tolerance. Evaluations using the public Cyber Autonomy Gym for Experimentation (CAGE) Challenge 2 environment demonstrate significant improvement over baselines that are oblivious to underlying human factors.**

***Index Terms*—cybersecurity threats, user behavior, adversarial inverse reinforcement learning, human factors.**

I. Introduction

Internet of Things (IoT) systems encompass a vast network of smart devices that communicate and interact over the Internet. Such IoT systems are inherently complex due to their integrative and interconnected nature, presenting unique and serious security challenges [1]–[3]. As the scale and complexity of IoT systems and wireless networks continue to rise, securing them against increasingly sophisticated attacks is becoming a challenging and strenuous task [4]. To this end, machine learning (ML) and artificial intelligence (AI) have demonstrated tremendous potential in automating cyber defense workflows and assisting operators in decision-making. The security defense in such complex networks actively involves human users, who often leverage their critical thinking and decision-making capabilities to augment and enhance the security and resilience of systems within IoT environments, known as Human-in-the-loop (HITL) [5].

The integration of an AI system, known as Assistive AI, can help model cyber operators' workflows and predict the next moves in cyber defense. This can be achieved through inverse learning of human-like policies using available behavioral data (e.g., observed human defense workflows and action sequences), enabling AI systems to provide recommendations to human operators, suggest corrections, and effectively support humans during operations. Despite the development of several AI and ML-based approaches for cyber defense, harnessing Assistive AI to improve the workflow of human operators by developing personalized models (e.g., reflecting different operating styles or expertise levels) to enhance the overall security posture of IoT networks has not been fully explored.

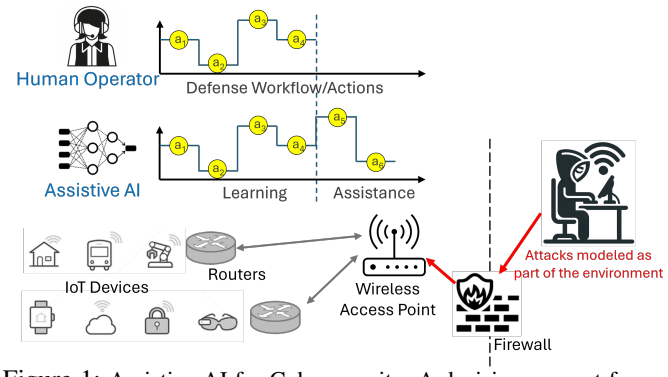

Figure 1: Assistive AI for Cybersecurity: A decision-support framework that leverages inverse reinforcement learning to provide action recommendations and behavior correction for human operators, enhancing cyber defense workflows in IoT environments.

This paper develops an adversarial inverse reinforcement learning framework that enables personalized workflow models conditioned on underlying human factors. The goal is to learn cyber defense policies for a wide range of cyber defense operators with varied human factors using available workflows and activity data. Imitation learning and inverse reinforcement learning (IRL) [6]–[10] learn policies and reward models of humans or experts using offline human data, e.g., expert demonstrations and trajectories. These methods learn policies or reward models that represent the average behavior of humans reflected in available demonstrations. Despite the success of existing inverse learning approaches in several domains [11], [12], human operators in cyber

defense vary significantly from each other. The behavioral differences are the latent factors that differentiate humans in terms of expertise, speed of response, risk tolerance, and more [13], [14]. The existing inverse learning algorithms that are oblivious to these human factors can only provide one-size-fits-all solutions, which limits their applicability in learning operator behaviors in cyber defense domains.

Specifically, we introduce a vector c to model the human factors differentiating human operators in cyber defense. Thus, the proposed approach learns a policy/model $\pi(a_t|s_t, c)$ that is conditioned on a human factor vector c. The factors could be a priori known (e.g., from surveys), or could be learned from the existing trajectories (e.g., using a classifier). To learn the policy/model, our proposed approach learns a discriminator, representing the user reward function, by contrasting the actions generated from the policy against demonstration trajectories. This framework enables us to generate personalized operator models/policies that better fit individual operators' strategies and behaviors, captured by human factors. The workflow of the Assistive AI is depicted in Fig. 1. This diagram illustrates a dynamic and interactive process where the AI system learns the human operator's policy and is able to predict the next human operator action sequence. More specifically, the actions labeled a_1 to a_4 are taken by a cyber defense operator, and the inverse learned model enables predicting the next action sequence of the operator, a_5 and a_6. The AI system with such capability can provide intelligent and context-aware suggestions and recommendations to the human operator, as well as assist the human operator against potential threats.

To evaluate the performance of the proposed method, we model users as suboptimal reinforcement learning agents. We consider three human factors in cyber defense: fatigue [15], expertise [13], and risk tolerance [16], [17]. These factors represent users with varied behaviors in cyber defense, influencing their speed, efficiency, and comfort to accept risk during the defense process. The performance of the proposed method is evaluated in the Cyber Autonomy Gym for Experimentation (CAGE) Challenge 2 environment [18]. The numerical experiments demonstrate significant improvement in learning accurate policies that capture varied behaviors for more effective IoT systems and wireless network defense reflecting underlying human factors.

This paper makes contributions in the realm of IoT and wireless network security. Firstly, it proposes a generative adversarial inverse reinforcement learning conditioned on human factors to learn user-specific policies/models aligned with human factors. Secondly, it introduces and models three distinct human factors that differentiate user/operator behavior in wireless cyber defense. Lastly, we demonstrate the efficacy and performance of the proposed method using the CAGE2 challenge environment, as an instance of wireless cyber defense scenarios.

II. Related Works

Several methods have been developed for the security of interconnected networks, given the importance of such networks in our infrastructure and daily lives. These include machine learning techniques for anomaly detection, robust authentication and access control to verify identities, secure offloading to manage resource constraints, and advanced malware detection to safeguard against cyber threats [1], [19], [20].

A wide range of these methods utilize deep learning techniques, ranging from real-time intrusion detection in wireless networks [21] to adversarial methods to ensure the robustness of security solutions [22].

Human-in-the-loop approaches further enhance this security framework by integrating human oversight into critical decision-making processes [5], [23]. This collaborative interplay is supported by research in learning from human or expert demonstrations, which has primarily focused on inverse reinforcement learning and imitation learning techniques [6], [10], [24]–[27]. IRL methods recover the expert's reward function from their behavior, while imitation learning directly learns the expert's policy via supervised learning. The close relationship between these methods is evident as recovering the reward function is akin to learning the policy, which is computable from the inferred reward function.

Recent advancements in deep neural networks have led to scalable IRL algorithms for both single-agent and multi-agent settings [28]. Meanwhile, Bayesian IRL offers a sample-efficient approach for learning through limited expert demonstrations [29], [30]. Furthermore, hierarchical IRL methods provide high-level guidance for complex problems, enhancing the causal relationship between different policy levels and redefined objectives for hierarchical policy learning [31], [32]. However, most existing methods learn policies that represent the average behaviors of humans reflected in demonstrations. Given the varied human operator behavior in cyberspace, these methods are unable to effectively recover policies that reflect cyber-defense users with varied human factors (i.e., behavior) using their user activity data.

III. Methodology

A. Problem Formulation for Network Defense Modeling

Human-in-the-loop defense mechanisms are essential in IoT security due to the domain knowledge and expertise of human operators to detect, defend against, and confirm attacks. Human operators' behavior in cyberspace differs from one another, such as varying expertise, risk tolerance, speed of response to threats, and more. Therefore, to accurately predict a human's behavior and next moves, it is essential to design assistive AI to learn individual human operator behavior and provide individualized support under various conditions.

We consider a wireless network consisting of n entities or components, each with distinct characteristics and contributing to the overall functionality of the network. An adversary – modeled as part of the environment in this paper – exploits vulnerabilities, compromises network entities, and propagates the attack/threats to other parts of the network. A network defender (i.e., an expert) makes a sequence of defense decisions to minimize the effect of adversarial attacks, such as integrity

monitoring, content analysis, credential hardening, execution isolation, and firewall updates.

Our objective is to learn the individualized behavior of users in cyberspace using some realizations of their behavioral data, τ_E. The data consists of state-action pairs made by various users, where user human-factor differences are denoted using a latent vector c. By utilizing user data and the underlying latent variables, our goal is to learn $\pi(a_t|s_t, c)$, which provides a range of policies across the latent space. The policy is conditioned on the human factors c, capturing the behavior of a specific human characterized by this factor. The human factors may be predetermined (e.g., derived from surveys) or inferred from the analysis of observed trajectories.

The user behavior and decision-making in cyber defense are modeled using a Markov decision process (MDP), expressed by a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, R \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the human action space, and $\mathcal{P}$ is the state transition probability function with $\mathcal{P}(\mathbf{s}, \mathbf{a}, \mathbf{s}') = p(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})$ expressing the probability that the next state will be $\mathbf{s}'$ if the user takes an action $\mathbf{a}$ in state $\mathbf{s}$. The transition probability is stochastic because adversaries are constantly aiming to compromise the networks, and their behavior is complex and unpredictable. Meanwhile, the attack propagation among the network components could also be stochastic, depending on the security level of the machines or other components of the network. The user reward function R quantifies the expected security gain from a user perspective, where $R(\mathbf{s}, \mathbf{a})$ denotes the expected reward earned when defense action $\mathbf{a}$ is taken by the user in state $\mathbf{s}$.

Considering the availability of the n components in the network, the state is a binary vector of size n, representing the status (compromised or not) for all n components. This vector is represented by $\mathbf{s}_k = [\mathbf{s}_k(1), ..., \mathbf{s}_k(n)]$, where $\mathbf{s}_k(i) = 1$ indicates that the ith component is compromised at the time step k and the reverse for $\mathbf{s}_k(i) = 0$. In particular, $\mathbf{s}_k = [0, 0, \cdots, 0]^T$ represents a network without any compromise, while $\mathbf{s}_k = [1, 1, \cdots, 1]^T$ represents a network with all nodes being compromised. Hence, the state vector can take 2^n different possible vectors, denoted by $\{\mathbf{s}^1, \cdots, \mathbf{s}^{2^n}\}$.

B. Inverse Learning with Human Factors

Inverse reinforcement learning is designed to deduce the underlying reward function from an expert's demonstrations. This inferred reward function can then be utilized to reconstruct the policy exhibited by the expert. Different human experts may exhibit varied behavioral patterns due to a range of inherent human factors [13], [14], [16]. Understanding the nuances of inherent human factors aids in crafting specifically tailored personalized models for various user types, aiming to enhance the performance of the IRL model. Generally, we can either infer the underlying human factor variables c of different trajectories using either a simple network (e.g., based on simple patterns and using supervised learning from known trajectories with human factor labels) or obtain the human factor variables by leveraging available side information (e.g., monitoring/sensor data and survey data). In this paper, we proceed with the assumption that human factors are either directly accessible or have been inferred from previous modeling work. We explore the benefits of integrating these factors into inverse reinforcement learning. In our study, we focus on a policy that is conditioned on a vector of human factors c, as depicted in Fig. 2.

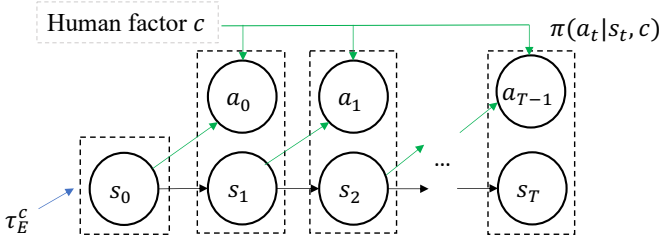

Figure 2: An illustration of our policy in inverse reinforcement learning, conditioned on human factors c.

We represent human attributes as a latent variable c. For example, different components of c might correspond to various aspects of human attributes. After c has been inferred, we are able to directly integrate it into the learning process. Specifically, we adopt the Maximum Entropy IRL framework, as proposed by [25], which treats IRL as a problem of maximum likelihood estimation (MLE), as formalized in Equation (2). In this context, instead of adopting a one-size-fits-all policy approach, we aim to learn a set of policies that are conditioned on the human factor variable c. This approach enables us to model diverse human behaviors and utilize data from various individuals to learn the set of policies. $\tau_E^c = (s_0^c, a_0^c, \ldots, s_T^c)$ signifies the expert's trajectory conditioned on the human factor c, which is a sequence of state and action pairs induced by the optimal policy $\pi^*(a|s, c)$ with a corresponding human factor. Previously, there was an interest in creating a unique model for each individual, but this effort was constrained by data scarcity, partly because available data from other individuals was not being utilized. Our approach addresses this issue by leveraging data from various sources to mitigate the data shortage. By conditioning on the human factor, we can develop personalized models tailored to fit the specific needs and characteristics of each person. Our primary objective is to learn the optimal policy $\pi^*(a|s, c)$ which will enable the construction of assistive AI agents through learning from human experts. Within the Maximum Entropy IRL formulation, the human-factor conditioned optimal policy $\pi^*(a|s, c)$ comes from the maximization of the expected entropy-regularized discounted reward, which is shown in Equation (1), where $H(\pi(\cdot|s_t))$ is the entropy regularization term to guide the policy to traverse the state space.

$$\pi^*(a|s, c) = \arg\max_{\pi} E_{\tau\sim\pi} \sum_{t=0}^{T} [\gamma^t (R_\theta(s_t, a_t|c) + H(\pi(\cdot|s_t))] \tag{1}$$

According to [33], the optimal policy follows the form $\pi^*(a|s, c) \propto exp(Q^*_{soft}(s_t, a_t|c))$, where $Q^*_{soft}(s_t, a_t|c) = R_\theta(s_t, a_t|c) + E_{(s_{t+1},\ldots)\sim\pi} \sum_{t'=t}^{T} [\gamma^{t'} (R_\theta(s_{t'}, a_{t'}|c) + H(\pi(\cdot|s_{t'}))]$ denotes the soft-Q function. The formulation of the MLE problem is shown in Equation (2), where $R_\theta(s_t, a_t|c)$

is the parameterization of the reward function, $p(s_{t+1}|s_t, a_t)$ is the state transition probability, $\mu(s_0)$ corresponds to the initial state distribution, and $\gamma \in (0, 1]$ is the discount factor in the infinite horizon case ($T = \infty$) to obtain a finite sum. The probability of a specific demonstration occurrence $\hat{P}_\theta(\tau_E^c)$ is proportional to the exponential of the corresponding sum of discounted reward, and $Z_\theta = \sum_{\tau_E^c} \hat{P}_\theta(\tau_E^c)$ represents the partition function for discrete cases, which is adopted to normalize the probabilities.

$$\max_\theta E_{\tau_E^c}[\log P_\theta(\tau_E^c)], P_\theta(\tau_E^c) = \frac{\hat{P}_\theta(\tau_E^c)}{Z_\theta}$$
$$\hat{P}_\theta(\tau_E^c) \propto \mu(s_0) \prod_{t=0}^{T-1} p(s_{t+1}|s_t, a_t) \exp(\gamma^t R_\theta(s_t, a_t|c)) \tag{2}$$

However, when dealing with large-scale state and action spaces, computing the partition function Z_θ becomes infeasible. To address this challenge, an adversarial approach (AIRL) was introduced by [34], which utilizes a sample-based method to approximate the MLE problem in a Generative adversarial network (GAN) framework [35]. $D_\theta(s_t, a_t|c) = \frac{\exp(f_\theta(s_t, a_t|c))}{\exp(f_\theta(s_t, a_t|c)) + \pi(a_t|s_t, c)}$ is the discriminator which involves alternate training of a discriminator network $f_\theta(s_t, a_t|c)$ and a policy $\pi(a_t|s_t, c)$. Different from [36], the discriminator $D_\theta^t = D_\theta(s_t, a_t|c)$ is with respect to a single state-action pair case instead of the whole trajectory. More precisely, the discriminator is updated through the minimization of the cross-entropy loss between the expert demonstrations τ_E^c and the policy-generated samples τ^c by $\pi(a|s, c)$: $\min_\theta \sum_{t=0}^{T-1}(-E_{\tau_E}[\log D_\theta^t] + E_\tau[\log(1 - D_\theta^t)])$. Simultaneously, the policy π is updated by maximizing the reward function formulated from the discriminator's outputs: $R(s_t, a_t|c) = \log D_\theta^t - \log(1 - D_\theta^t)$ using Q-learning. Within this Q-learning framework, the policy π is acquired by $\pi(a|s, c) = \arg\max_a[\frac{1}{Z_s} \exp(Q(s, a|c))]$, where $Z_s = \sum_{a'} \exp(Q(s, a'|c))$ serves as a normalization factor. Upon reaching optimality, the discriminator $f_\theta(s, a|c)$ effectively acts as the reconstructed reward function $R_\theta(s, a|c)$, while the optimal policy $\pi^*(a|s, c)$ emerges as the replicated expert policy. The proposed algorithm is summarized in Algorithm 1.

C. Modeling Human Factors

This paper introduces three human factors that differentiate users' behavior in cyber defense. The user reward function, R, governs how a user acts under different conditions (e.g., responses to different threats). The reward function representing the primary objective for all users can be factorized in terms of security and action reward as:

$$R = R^{\text{Security}} - C^{\text{Action}}, \tag{3}$$

where the first term, R^{Security}, represents the reward with respect to network security, and the second term, C^{Action}, indicates the cost of defense actions.

This paper models the following three human factors for cyber defense users:

Fatigue: This factor represents the speed of the user's response to threats. In this paper, we model the fatigue level with the parameter $\zeta \in [0, 1]$, where $\zeta = 1$ represents fast-acting users, whereas a smaller ζ indicates a user who may miss acting on time with a probability of $1 - \zeta$. A sample from this random variable. i.e., $z_k \sim \text{Bernoulli}(\zeta)$, models a possible reaction of the user during the defense process. $z_k = 1$ represents the case where a user takes action, whereas $z_k = 0$ indicates the case where the user misses the action at time step k. The user's reward is also impacted by this sample as follows:

$$R = R^{\text{Security}} - z_k C^{\text{Action}}, \tag{4}$$

Risk Tolerance: This factor plays a pivotal role in differentiating user behavior and response strategies. Risk tolerance is essentially the degree to which a user is willing to engage with potential threats within the network. Individuals with low risk tolerance are typically more proactive and aggressive in their response to threats [16], [17].

$$R = R^{\text{Security}} - \beta C^{\text{Action}}, \tag{5}$$

where larger β models a higher cost for actions, resulting in a more risk-tolerant user.

Expertise: This refers to the degree of proficiency and knowledge that users possess in defending against threats. We model the expertise level of users using a parameter $\alpha \in [0, 1]$, where α represents the probability that the user's actions are effective in eliminating network compromises. A value of α near 1 represents skilled users, while an α closer to 0 indicates users with little to no expertise in removing compromises.

According to the three human factors discussed above, each human behavior can be characterized using a vector $c = [\zeta, \beta, \alpha]$, where the elements measure a single human behavior across different factors.

IV. Computational Experimentation

A. Experiment Setup

We utilize the public CAGE Challenge 2 environment for our experiments [18]. The CAGE 2 challenge is an initiative aimed at advancing autonomous cyber operations (ACO), which are critical in the dynamic field of cybersecurity. The CAGE Challenge 2 scenario is a manufacturing plant network that is under attack by a neighboring nation. The goal of the challenge is to create an autonomous defender that can defend a network against adversaries that are trying to disrupt network operations and access critical information in servers. The attackers/adversaries are referred to as red agents, and the defenders are defined as blue agents.

The network structure is shown in Fig. 3, which consists of three sub-networks. The status of network components (compromised or not) has been represented using a binary state vector of size 16, where 1 and 0 at each variable indicate the compromise or lack of compromise, respectively, for specific components of the network. The attack propagation in this network system unfolds through a sequence of conditional activations based on the current compromise status of various components. The attack penetrates the network through user hosts in subnet 1 and propagates in the network.

Algorithm 1 Adversarial Inverse Reinforcement Learning with Human Factors

Input: Expert demonstrations $\tau_E^c = (s_0^c, a_0^c, \ldots, s_T^c)$ with human factors c.

Initialize the discriminator network f_θ and the policy π.

for *each training episode* **do**

  Generate M trajectories $\{\tau_E^c\}$ by interacting with the simulator.

  Update $f_\theta(s_t, a_t|c)$ by minimizing the cross entropy loss: $\sum_{t=0}^{T-1}(-E_{\tau_E}[\log D_\theta^t] + E_\tau[\log(1 - D_\theta^t)])$.

  Update π based on the generated trajectories $\{\tau_E^c\}$ and corresponding rewards $R_t = \log D_\theta^t - \log(1 - D_\theta^t)$ using Q-Learning method.

end for

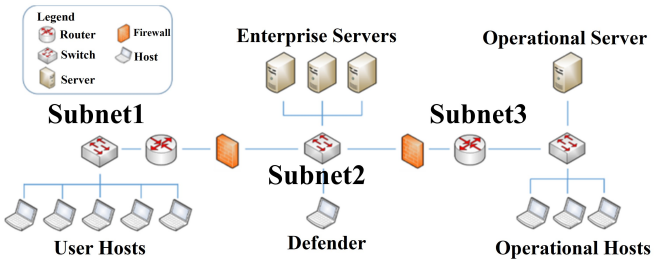

Figure 3: An illustration of the CAGE2 challenge environment.

The adversarial and defense strategies in this scenario involve tactical alterations in the compromised status of user hosts within subnet 1. The adversary (or red agent) executes actions to alter the status of five user hosts, denoted by $\mathbf{s}(1)$ to $\mathbf{s}(5)$, from an uncompromised state (0) to a compromised state (1). In response, the defense agent (e.g., user) undertakes defensive actions by re-imaging or restoring the affected user hosts in subnet 1.

In the CAGE2 challenge, compromise at different components of the network causes different security risks. When the attacker accesses user hosts, i.e., $\mathbf{s}(1)$ to $\mathbf{s}(5)$, the security reward of -0.1 is considered for each of the compromised user hosts. The stakes are higher when the attacker breaches enterprise servers $\mathbf{s}(8)$, $\mathbf{s}(9)$, $\mathbf{s}(10)$, with each incident resulting in a -1 reward. The most critical scenario is the compromise of the operational server $\mathbf{s}(16)$ by the attacker, leading to a substantial -10 reward.[1] The reward of each restoring or reimaging as part of defense actions is assumed to be -1, which ensures the highest security with minimum allocated resources (i.e., costs and/or disruption in the network operations).

The user defends the network against adversarial actions without knowledge of the adversary's policy. The user policy is modeled using an imperfect reinforcement learning agent, which represents the stochastic form of the optimal policy in (1) according to ϵ-greedy model with $\epsilon = 0.1$. Due to the finite state and action spaces, the adversarial and user policies are obtained using the Q-Learning algorithm [37].[2] Note that separate Q-Learning algorithms are trained to represent user policies associated with different human factors, as each factor represents different rewards and/or transition probabilities.

B. Performance of AIRL with Human Factors

In our analysis, we considered varying values for the three human factors: fatigue, risk tolerance, and expertise. The impact of these human factors on defense process performance is analyzed and shown in Fig. 4. The left plot shows that less fatigued users ($\zeta = 1$) achieve higher network security through fast, timely defense actions. In contrast, highly fatigued users ($\zeta = 0.2$) react slowly, leading to lower security. The middle plot indicates that users with higher expertise ($\alpha = 1$) maintain better network security, while those with lower skills ($\alpha = 0.2$) struggle against threats, resulting in quicker attack progression and reduced security. The right plot reveals that proactive, less risk-tolerant users ($\beta = 0.2$) attain better security. In contrast, risk-tolerant users ($\beta = 5$) engage minimally in defense, leading to lower security. Fig. 5 compares the number of restored user hosts between a risk-averse (proactive) user ($\beta = 1$) and a risk-tolerant user ($\beta = 5$), showing more proactive restoration by the risk-averse user.

To illustrate the effects of human factors on recovered rewards, we present a comparative analysis in Fig. 6, where each human factor is individually examined. Specifically, we adjusted one of the three factors — fatigue, risk tolerance, or expertise — to a value of 0.2, while maintaining the other two at a constant value of 1 for reference.

The purple line with $c_1 = [0.2, 1, 1]$, indicative of high-risk aversion, attained the highest recovered reward, suggesting a strategy that effectively emphasizes cautious decision-making. Besides, the green line with $c_2 = [1, 0.2, 1]$, while slightly lower, still presents a comparable reward, indicating that higher fatigue does not drastically impair the performance of the user. On the other hand, the red line with $c_3 = [1, 1, 0.2]$, representing less expertise, resulted in the lowest reward, highlighting the adverse impact that lower skill level has on the efficacy of cyber defense operations.

To illustrate the impact of incorporating human factor information into our AIRL model, we conducted a comparative

[1] In the context of reinforcement learning, the cost of actions is represented as a negative reward, reflecting the effort or resources required to perform the actions.

[2] The adversary is assumed to be part of the environment and the Q-learning algorithm is used to obtain the defense policy.

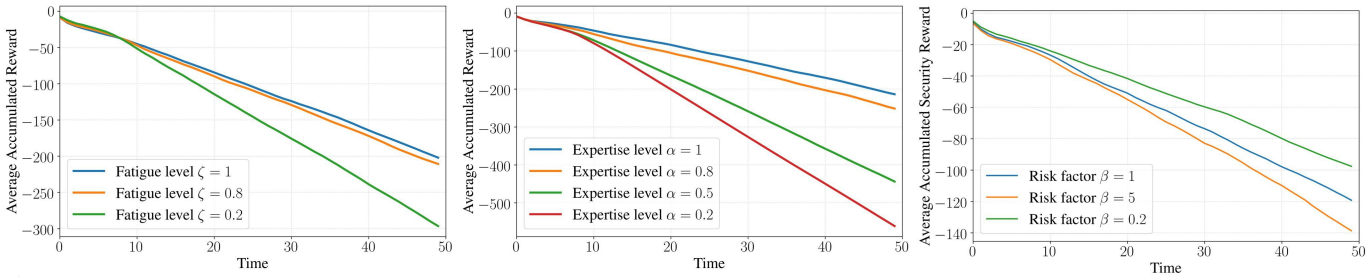

Figure 4: Left plot: Average accumulated reward for users with different fatigue levels. Middle plot: Average accumulated reward for users with different expertise levels. Right plot: Average accumulated security reward for users with different risk levels.

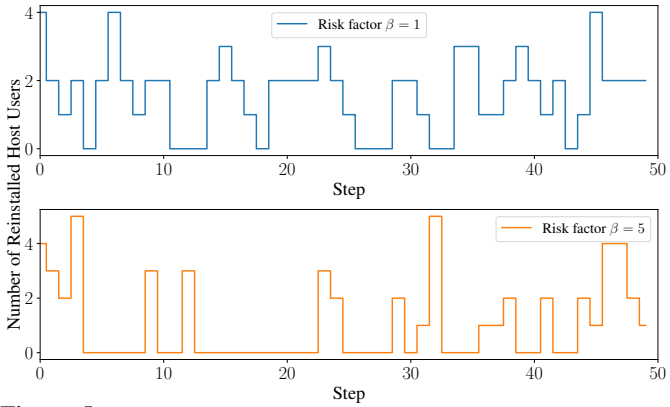

Figure 5: Comparison of defense actions over time taken by risk-averse (proactive) (β = 1) and risk-tolerant (β = 5) users.

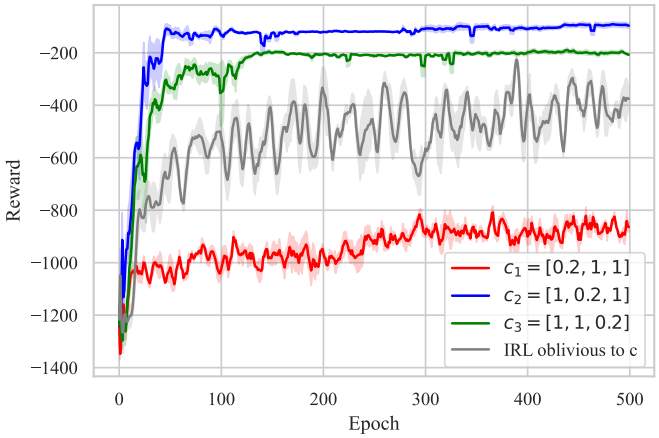

Figure 6: Average recovered rewards for users with different human factors.

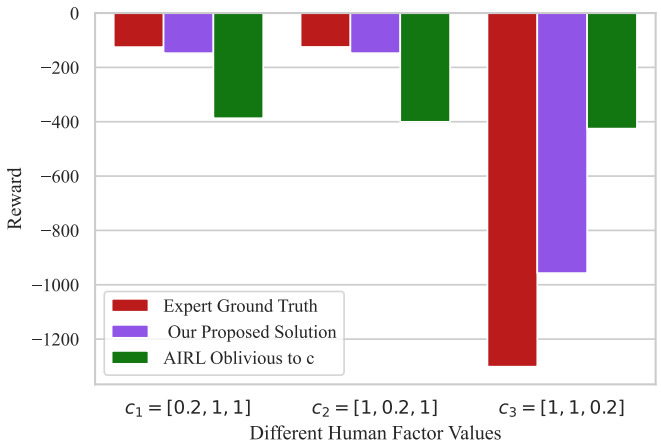

Figure 7: Comparison of the average recovered reward between the proposed method and the baseline IRL policy oblivious to human factors.

analysis of AIRL performance both with and without the inclusion of human factors. In scenarios where human factors are not integrated into our AIRL framework, the model learns a universal approach, disregarding individual differences among users. This generalized, one-size-fits-all strategy fails to offer customized models tailored to the unique characteristics of different human operators in cybersecurity. For our experiment, we randomly selected each element of the latent human factors $c = [\zeta, \beta, \alpha]$ from a uniform distribution between 0 and 1 and subsequently generated expert demonstrations based on these factors. When mixing trajectories from diverse c values without disclosing them to the AIRL model, a generalized policy is learned. However, by incorporating human factors and making them visible alongside expert demonstrations, we can facilitate personalized IRL for distinct groups of defense users, thereby achieving more targeted and effective performance.

Fig. 6 shows the convergence of our AIRL model conditioned on different human factors c_1, c_2, and c_3 with a comparison with the old model that is oblivious to the human factor c. The colored curves corresponding to different types of humans converge to different recovered rewards quickly. However, the grey line, which does not take the human factor into account, converges slowly. The graph shows that including the human factor information helps to develop a personalized model and makes inverse reinforcement learning faster.

To provide a detailed comparison of the rewards obtained under different human factor settings after the Adversarial Inverse Reinforcement Learning (AIRL) algorithm stabilizes, we executed the training of AIRL over 1,000 epochs for each scenario. Fig. 7 displays a bar chart depicting the average reward ascertained through AIRL, with consideration of human factor c, across the last 100 epochs, shown in purple. For context, we juxtaposed this with the average reward from expert demonstrations, highlighted in red, and the average reward obtained by AIRL without considering human

factors, shown in green. Across various groups characterized by human factors c_1, c_2, and c_3, our approach more accurately approximates the true reward compared to the previous method that disregards human factor c. This indicates that incorporating human factor information significantly enhances our model's ability to perform inverse learning tailored to different individuals.

The analysis presented in Fig. 4 highlights that the level of expertise is the most significant human factor, leading to the greatest variation in rewards. In Fig. 8, we explored the impact of expertise to the recovered reward by varying the expertise factor from 0.1 to 1 and fix the fatigue and risk factors to different combinations, with other settings same as Fig. 6. The findings reveal that rewards increase with the expertise factor, although this growth diminishes at higher levels of expertise. This suggests that the additional benefit of increasing expertise diminishes when individuals are already highly skilled, with rewards beyond a certain threshold. This observation underscores the importance of ensuring that all defense agents meet a minimum level of expertise to maintain reliable defense performance. Furthermore, the gaps between different curves highlights the relative influence of various human factors. The proximity of the orange curve to the default blue line showing no fatigue and neutral risk-tolerant indicates that the fatigue factor has a minimal impact. Conversely, the green curve's position above the baseline and the significant drop of the red line below it confirm that individuals with a higher tolerance for risk tend to secure higher rewards, aligning with common expectations.

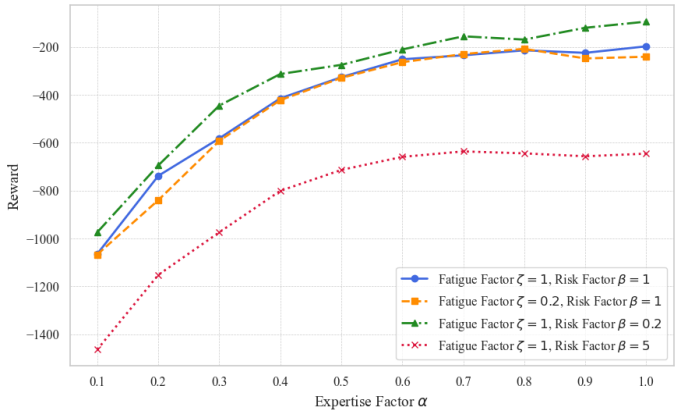

Figure 8: Average recovered rewards for users with different expertise levels.

V. Conclusions and Future Work

This paper introduces an adversarial inverse reinforcement learning approach to recover the user policy in IoT systems and wireless cyber defense. The variation in user behavior is considered by introducing three human factors: expertise, fatigue, and risk tolerance. These factors are mathematically modeled in terms of the user reward function in the reinforcement learning context. The proposed method takes data from users with different characteristics (human factors) into account and learns policies specifically tailored to specific users (i.e., users with specific human factors). The proposed method enhances the accurate understanding/prediction of user behavior, helping to design AI security solutions to provide the best user support in cyberspace. Our method has been tested on the CAGE2 environment. The numerical results demonstrate significant variations in the behaviors of users with different human factors and the importance of learning user-specific policies to predict user behavior. The future work consists of modeling the decoy behavior of the attacker and finding sophisticated strategies by the attacker in terms of Markov games.

Acknowledgements

This work was supported in part by the U.S. Military Academy (USMA) under Cooperative Agreement No. W911NF-23-2-0175 and the OUSD(Research & Engineering) Cyber Technologies Office under Support Agreement No. USMA 21057. The views and conclusions expressed in this paper are those of the authors and do not reflect the official policy or position of the U.S. Military Academy, U.S. Army, U.S. Department of Defense, or U.S. Government.

References

[1] M. A. Al-Garadi, A. Mohamed, A. K. Al-Ali, X. Du, I. Ali, and M. Guizani, "A survey of machine and deep learning methods for internet of things (iot) security," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 3, pp. 1646–1685, 2020.

[2] A. Kazeminajafabadi and M. Imani, "Optimal monitoring and attack detection of networks modeled by Bayesian attack graphs," *Cybersecurity*, vol. 6, no. 1, p. 22, 2023.

[3] R. Zandi, K. Behzad, E. Motamedi, H. Salehinejad, and M. Siami, "Robofisense: Attention-based robotic arm activity recognition with wifi sensing," *IEEE Journal of Selected Topics in Signal Processing*, 2024.

[4] A. Kazeminajafabadi and M. Imani, "Optimal joint defense and monitoring for networks security under uncertainty: A POMDP-based approach," *IET Information Security*, vol. 2024, no. 1, p. 7966713, 2024.

[5] M. Tyworth, N. A. Giacobe, V. F. Mancuso, M. D. McNeese, and D. L. Hall, "A human-in-the-loop approach to understanding situation awareness in cyber defence analysis," *EAI Endorsed Transactions on Security and Safety*, vol. 1, no. 2, 2013.

[6] J. Ho and S. Ermon, "Generative adversarial imitation learning," *Advances in neural information processing systems*, vol. 29, 2016.

[7] X. Zhang, K. Zhang, E. Miehling, and T. Basar, "Non-cooperative inverse reinforcement learning," *Advances in neural information processing systems*, vol. 32, 2019.

[8] A. Ravari, S. F. Ghoreishi, and M. Imani, "Optimal inference of hidden Markov models through expert-acquired data," *IEEE Transactions on Artificial Intelligence*, 2024.

[9] Y. Lin, S. F. Ghoreishi, T. Lan, and M. Imani, "High-level human intention learning for cooperative decision-making," in *2024 IEEE Conference on Control Technology and Applications (CCTA)*, pp. 209–216, IEEE, 2024.

[10] J. Song, H. Ren, D. Sadigh, and S. Ermon, "Multi-agent generative adversarial imitation learning," *Advances in neural information processing systems*, vol. 31, 2018.

[11] W. Liu, J. Zhong, R. Wu, B. L. Fylstra, J. Si, and H. H. Huang, "Inferring human-robot performance objectives during locomotion using inverse reinforcement learning and inverse optimal control," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2549–2556, 2022.

[12] B. Lian, W. Xue, F. L. Lewis, and T. Chai, "Inverse reinforcement learning for adversarial apprentice games," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[13] T. Rahman, R. Rohan, D. Pal, and P. Kanthamanon, "Human factors in cybersecurity: a scoping review," in *The 12th International Conference on Advances in Information Technology*, pp. 1–11, 2021.

[14] J. Jeong, J. Mihelcic, G. Oliver, and C. Rudolph, "Towards an improved understanding of human factors in cybersecurity," in *2019 IEEE 5th International Conference on Collaboration and Internet Computing (CIC)*, pp. 338–345, IEEE, 2019.

[15] C. Nobles, "Stress, burnout, and security fatigue in cybersecurity: A human factors problem," *HOLISTICA–Journal of Business and Public Administration*, vol. 13, no. 1, pp. 49–72, 2022.

[16] V. Dutt, Y.-S. Ahn, and C. Gonzalez, "Cyber situation awareness: Modeling the security analyst in a cyber-attack scenario through instance-based learning," in *Data and Applications Security and Privacy XXV: 25th Annual IFIP WG 11.3 Conference, DBSec 2011, Richmond, VA, USA, July 11-13, 2011. Proceedings 25*, pp. 280–292, Springer, 2011.

[17] X. Wu and H. Liao, "A compensatory value function for modeling risk tolerance and criteria interactions in preference disaggregation," *Omega*, vol. 117, p. 102836, 2023.

[18] CAGE, "Ttcp cage challenge 2." https://github.com/cage-challenge/cage-challenge-2, 2022.

[19] L. Xiao, X. Wan, X. Lu, Y. Zhang, and D. Wu, "Iot security techniques based on machine learning: How do iot devices use ai to enhance security?," *IEEE Signal Processing Magazine*, vol. 35, no. 5, pp. 41–49, 2018.

[20] R. Mahmoud, T. Yousuf, F. Aloul, and I. Zualkernan, "Internet of things (iot) security: Current status, challenges and prospective measures," in *2015 10th international conference for internet technology and secured transactions (ICITST)*, pp. 336–341, IEEE, 2015.

[21] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 375–388, 2020.

[22] F. Restuccia, S. D'Oro, A. Al-Shawabka, B. C. Rendon, K. Chowdhury, S. Ioannidis, and T. Melodia, "Generalized wireless adversarial deep learning," in *Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning*, pp. 49–54, 2020.

[23] D. S. Nunes, P. Zhang, and J. S. Silva, "A survey on human-in-the-loop applications towards an internet of all," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 2, pp. 944–965, 2015.

[24] A. Y. Ng, S. Russell, *et al.*, "Algorithms for inverse reinforcement learning.," in *Icml*, vol. 1, p. 2, 2000.

[25] B. D. Ziebart, A. L. Maas, J. A. Bagnell, A. K. Dey, *et al.*, "Maximum entropy inverse reinforcement learning.," in *Aaai*, vol. 8, pp. 1433–1438, Chicago, IL, USA, 2008.

[26] Z. Wu, L. Sun, W. Zhan, C. Yang, and M. Tomizuka, "Efficient sampling-based maximum entropy inverse reinforcement learning with application to autonomous driving," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5355–5362, 2020.

[27] S. Schaal, "Learning from demonstration," *Advances in neural information processing systems*, vol. 9, 1996.

[28] L. Yu, J. Song, and S. Ermon, "Multi-agent adversarial inverse reinforcement learning," in *International Conference on Machine Learning*, pp. 7194–7201, PMLR, 2019.

[29] D. Ramachandran and E. Amir, "Bayesian inverse reinforcement learning.," in *IJCAI*, vol. 7, pp. 2586–2591, 2007.

[30] S. Balakrishnan, Q. P. Nguyen, B. K. H. Low, and H. Soh, "Efficient exploration of reward functions in inverse reinforcement learning via bayesian optimization," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4187–4198, 2020.

[31] J. Chen, T. Lan, and V. Aggarwal, "Hierarchical adversarial inverse reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[32] J. Zhang, H. Yu, and W. Xu, "Hierarchical reinforcement learning by discovering intrinsic options," *arXiv preprint arXiv:2101.06521*, 2021.

[33] B. D. Ziebart, *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. Carnegie Mellon University, 2010.

[34] J. Fu, K. Luo, and S. Levine, "Learning robust rewards with adversarial inverse reinforcement learning," *arXiv preprint arXiv:1710.11248*, 2017.

[35] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[36] C. Finn, P. Christiano, P. Abbeel, and S. Levine, "A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models," *arXiv preprint arXiv:1611.03852*, 2016.

[37] J. Clifton and E. Laber, "Q-learning: Theory and applications," *Annual Review of Statistics and Its Application*, vol. 7, pp. 279–301, 2020.