engineering proceedings

MDPI

Proceeding Paper

Reward Hacking in Reinforcement Learning-Based Myoelectric Prosthesis Control: Challenges Beyond Tracking-Based Rewards †

Mario Moreno-Pallares *, Lorena Isabel Barona López, Ángel Leonardo Valdivieso Caraguay and Marco E. Benalcázar

Artificial Intelligence and Computer Vision Research Lab, Departamento de Informática y Ciencias de la Computación (DICC), Escuela Politécnica Nacional, Quito 170143, Ecuador; lorena.barona@epn.edu.ec (L.I.B.L.); angel.valdivieso@epn.edu.ec (Á.L.V.C.); marco.benalcazar@epn.edu.ec (M.E.B.)

* Correspondence: mario.moreno01@epn.edu.ec

† Presented at the Sixth International Conference on Information Systems and Software Technologies (ICI2ST 2026), Quito, Ecuador, 23–25 September 2026.

Abstract

Reward function design is a critical challenge in reinforcement learning (RL)-based myoelectric prosthesis control. Many existing approaches rely on tracking-based reward formulations that assume access to target trajectories or reference information during both training and deployment. However, such information is unavailable in practical amputee scenarios, potentially introducing hidden dependencies between the reward function and the observation space. This study investigates the emergence of reward hacking when transitioning from tracking-based to target-free observation spaces in a 4-DoF myoelectric prosthesis controlled using Double Deep Q-Networks and an 81-action discrete action space. Experimental results show that reward functions originally designed for reference-guided control tasks can induce degenerate policies characterized by persistent null actions and reduced motor activity. To address this issue, a modified reward formulation was developed to encourage exploration and active control behaviors. The proposed approach reduced action collapse and improved action diversity. These findings highlight the importance of jointly designing reward functions and observation spaces for deployable RL-based prosthetic controllers.

Keywords: reward hacking; reinforcement learning; myoelectric prosthesis; prosthetic hand control; reward function design; observation space

1. Introduction

Upper-limb amputation significantly affects an individual's ability to perform activities of daily living, motivating continuous advances in intelligent prosthetic technologies and human–machine interfaces. Among the available approaches, myoelectric prostheses have become one of the most promising solutions due to their ability to decode user intention from surface electromyographic (sEMG) signals and translate it into prosthetic movements [1,2].

Traditional myoelectric control strategies commonly rely on pattern recognition and supervised learning methods to classify predefined movement intentions. Although these approaches have achieved considerable success in laboratory environments, they often suffer from limited adaptability, reduced robustness to signal variability, electrode displacement, and changes in muscle activation patterns over time [2,3]. These limitations become more pronounced as the number of controllable degrees of freedom increases.

Check for updates

Academic Editor: Samuel Lascano

Published: 18 September 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

RL has recently emerged as an alternative for prosthetic control because it enables agents to learn control policies through interaction with the environment rather than relying exclusively on labeled training data. Several studies have demonstrated the potential of RL for improving adaptation, personalization, and autonomous tuning in assistive robotic systems and prosthetic devices [4,5]. More recently, deep RL approaches have been investigated for grasp optimization and prosthetic hand actuation tasks involving multiple degrees of freedom [6].

Despite these advances, reward function design remains one of the least standardized and most challenging aspects of RL-based prosthetic control [7]. Many reinforcement learning studies for myoelectric prosthesis control are developed and evaluated in simulated environments where target trajectories, desired joint positions, or reference kinematic signals are explicitly available during training. Consequently, reward formulations are naturally expressed in terms of trajectory tracking errors or deviations from predefined reference movements. Although this approach facilitates learning and benchmarking, it implicitly assumes the availability of information that is typically inaccessible during practical prosthesis operation.

From a RL perspective, these variables constitute privileged information, since they simplify the learning problem without being available during deployment [8]. As a result, reward functions that perform successfully under tracking-oriented conditions may become inconsistent when the agent is required to operate using only the sensory information available in real myoelectric prostheses [9,10]. Despite its practical importance, the interaction between privileged information, reward design, and observation space has received limited attention in reinforcement learning-based prosthetic control.

However, this assumption is problematic in real-world amputee applications. During practical operation, amputee users do not possess an accessible ground-truth representation of the intended prosthetic motion, and therefore target trajectories or reference errors cannot be directly measured or provided to the controller [11,12]. Consequently, observation spaces that include reference trajectories or tracking errors may introduce privileged information that artificially simplifies the learning problem and overestimates real deployment performance [9,10].

The removal of such privileged information can expose hidden dependencies between the reward function and the observation space. In these scenarios, agents may discover unintended strategies that maximize the numerical reward while failing to achieve the actual control objective. This phenomenon, commonly referred to as reward hacking or specification gaming, has become an important topic in artificial intelligence (AI) safety and alignment research [9,10,13]. Rather than learning the desired behavior, the agent exploits loopholes or ambiguities in the reward specification to obtain high returns through undesirable or degenerate policies.

Although reward hacking has been extensively discussed in general RL literature, its implications for myoelectric prosthesis control remain largely unexplored. This issue is particularly relevant in prosthetic applications because controllers are expected to operate under severe sensing limitations and incomplete information regarding user intentions.

Therefore, the present study investigates the emergence of reward hacking during the transition from tracking-oriented observation spaces to target-free observation spaces intended for deployment in practical myoelectric prosthesis controllers. The analysis is conducted using the Handi EPN V2 prosthetic control platform, a four-degree-of-freedom (4-DoF) myoelectric prosthetic hand controlled through a Double Deep Q-Network (DDQN) agent operating over an 81-action discrete action space. Particular attention is given to the behavioral changes that occur after removing privileged target information from the observation space while maintaining the same RL algorithm and tracking-based reward

https://doi.org/10.3390/engproc2026153005

formulation. This experimental setting enables the identification of hidden dependencies between reward functions and observation spaces, as well as the evaluation of a redesigned reward formulation specifically developed for target-free operation.

The central hypothesis of this study is that reward hacking in RL-based myoelectric prosthesis control is not necessarily caused by an incorrect reward formulation itself, but by a reward–observation mismatch, i.e., an incompatibility between the reward function and the information available within the observation space. Under this hypothesis, a reward designed for tracking-oriented learning may become vulnerable after privileged target information is removed, even if the RL algorithm remains unchanged. Consequently, reward functions should be evaluated together with the observation spaces in which they are intended to operate rather than as independent design components.

This hypothesis is experimentally evaluated through a controlled comparison of three experimental configurations that differ only in their observation space and reward formulation while maintaining the same reinforcement learning algorithm, action space, training protocol, and evaluation procedure. This experimental design enables the effect of reward–observation consistency to be isolated from other sources of variability.

The principal contributions of this study are summarized as follows:

- It is demonstrated that reward–observation mismatch can emerge solely as a consequence of removing privileged target information from the observation space, even when the reward formulation and reinforcement learning algorithm remain unchanged. This finding reveals a previously unexplored dependency between reward design and state representation in RL-based myoelectric prosthesis control.
- A set of behavioral metrics, including the Null Action Ratio (NAR), Effective Action Space Usage (EASU), Action Entropy, and Average Joint Velocity (AJV), is introduced to identify policy collapse and inactive behaviors that cannot be reliably detected using conventional tracking metrics alone.
- A reward formulation specifically designed for target-free prosthetic control is presented, demonstrating that inactive policy collapse can be significantly mitigated by aligning the reward structure with the information realistically available during deployment.
- It is shown that observation spaces and reward functions should be jointly designed rather than independently optimized. This principle extends beyond myoelectric prosthesis control and provides practical guidelines for developing deployable RL systems operating without privileged information.

2. Materials and Methods

The overall experimental workflow adopted in this study is illustrated in Figure 1. The diagram summarizes the interaction between the myoelectric prosthetic environment, the reinforcement learning agent, the observation space, the discrete action space, and the reward function. The subsequent subsections describe each component of the framework in detail.

2.1. Experimental Platform

The experiments were conducted using the Handi EPN V2 myoelectric prosthetic control platform, an experimental framework developed to investigate RL algorithms for upper-limb prosthesis control. The platform integrates sEMG signal acquisition, a 4-DoF prosthetic hand model, a RL environment implemented in MATLAB R2023b, and evaluation tools for behavioral and control performance analysis.

https://doi.org/10.3390/engproc2026153005

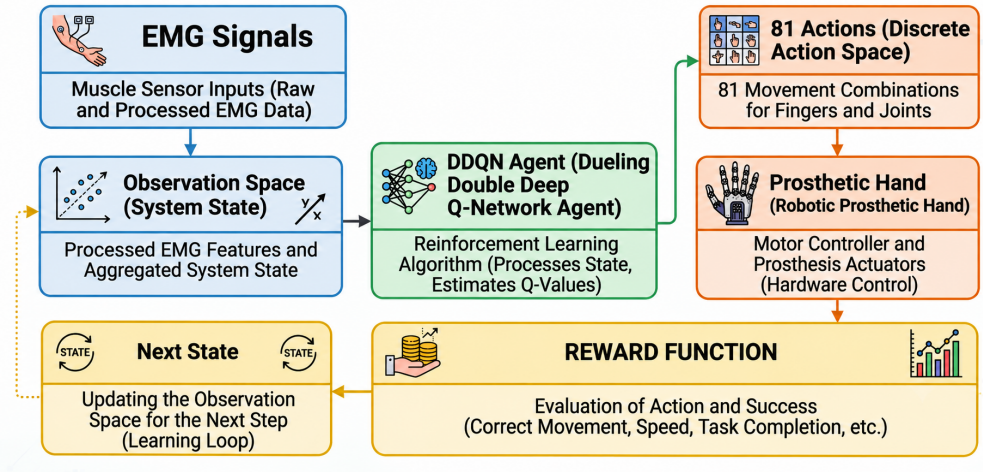

Figure 1. Overall experimental workflow of the proposed RL framework.

The implementation of the prosthetic environment and associated control framework is publicly available through the Handi EPN V2 repository developed at Escuela Politécnica Nacional (EPN) for RL-based prosthetic research [14].

The Handi EPN V2 platform serves as the experimental testbed used throughout this study and should not be interpreted as a contribution of the present work. Instead, the scientific contribution focuses on the interaction between reward formulation and observation space during RL.

The control problem was formulated as a discrete-time RL task in which the agent interacted with the environment by observing the current prosthetic state and selecting actions from a predefined discrete action space. Following each action execution, the environment updated the prosthetic state and returned both the subsequent observation and a scalar reward value computed according to the reward formulation under evaluation.

The control loop operated with a sampling period of 0.1 s and a maximum episode length of 100 interaction steps, corresponding to a maximum duration of 10 s. During evaluation with prerecorded signals, an episode could terminate earlier when the available EMG and glove sequence was exhausted; therefore, the effective number of interaction steps was recorded for every episode.

The experimental environment was specifically designed to analyze the transition from reference-guided control scenarios, where privileged target information is available, to target-free operation intended for practical amputee deployment. This configuration enabled the study of hidden dependencies between reward formulations and observation spaces, as well as the emergence of reward hacking behaviors under incomplete information conditions.

Figure 2 summarizes the physical implementation of the Handi EPN V2 experimental platform employed throughout this work. The platform integrates the MYO armband (Thalmic Labs Inc., Kitchener, ON, Canada) for surface electromyographic signal acquisition, a CyberGlove III developed at Escuela Politécnica Nacional (EPN, Quito, Ecuador) used to acquire reference hand kinematics during tracking-based experiments, the four-degree-of-freedom Handi EPN V2 prosthetic hand developed at EPN (Quito, Ecuador), the embedded motor control electronics developed at EPN (Quito, Ecuador), and the host computer executing the reinforcement learning environment developed in MATLAB R2023b. This configuration enables both tracking-based experiments and target-free evaluations under identical hardware conditions.

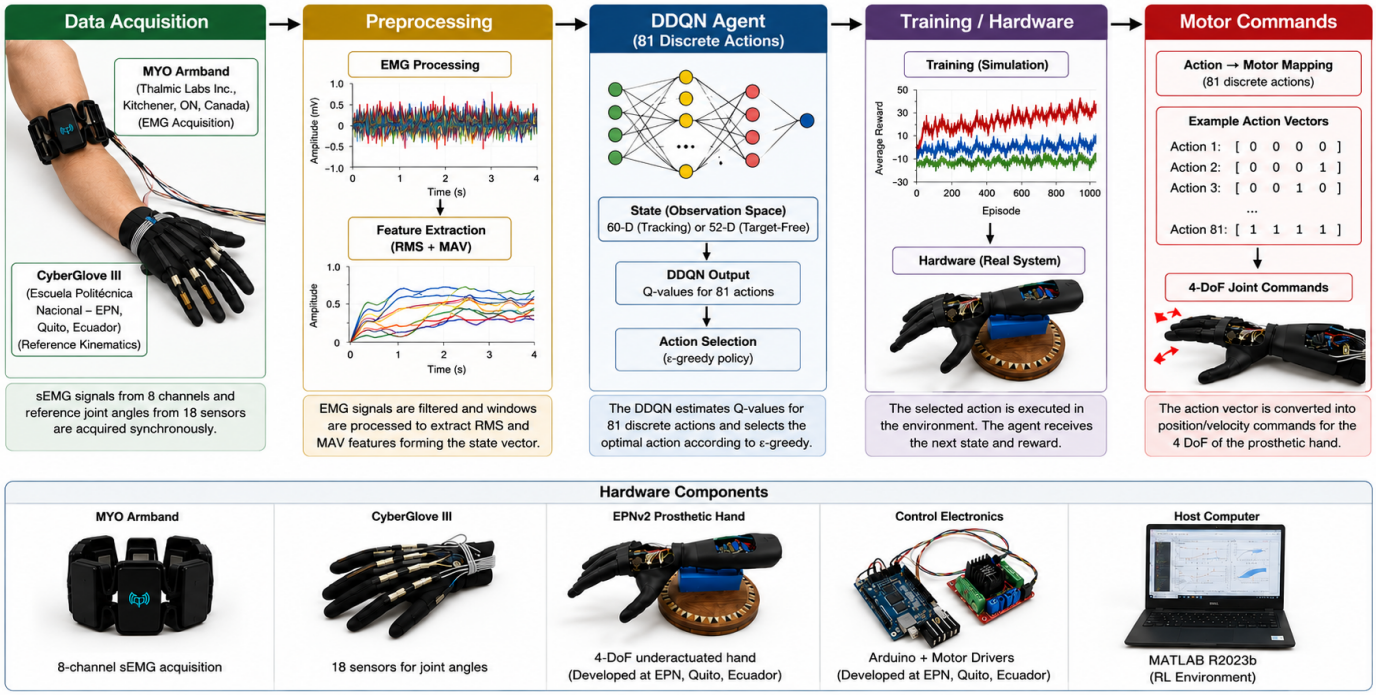

Figure 2. Overview of the Handi EPN V2 experimental platform. The system integrates the EPN V2 prosthetic hand, MYO armband, CyberGlove III, embedded control electronics, and host computer used for reinforcement learning training and evaluation.

The overall interaction between the observation space, the DDQN agent, the prosthetic environment, and the reward computation process is illustrated in Figure 3.

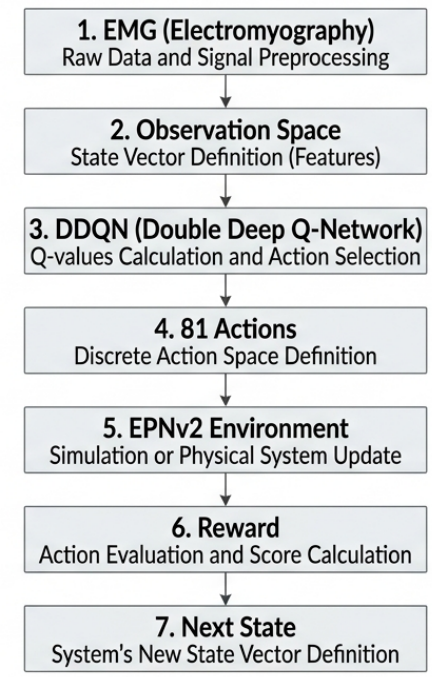

Figure 3. Learning framework employed in the Handi EPN V2.

2.2. Observation Spaces

Two observation spaces were investigated in this study to analyze the relationship between state representation and reward design in RL-based prosthetic control.

The first observation space, denoted as State60, corresponded to a tracking-oriented configuration that included sEMG features, current joint positions, predicted target positions, tracking errors, joint velocities, and the previously executed action:

$$\text{State60} = [\text{EMG}, q, q_{\text{ref}}, e, \dot{q}, a_{t-1}], \quad (1)$$

where EMG represents the vector of processed surface electromyographic features acquired from the user, q denotes the current prosthetic joint positions, q_{ref} denotes the target joint positions, $e = q_{\text{ref}} - q$ is the tracking error, $\dot{q}$ represents the joint angular velocities of the prosthetic hand, and a_{t-1} corresponds to the action applied during the previous interaction step.

This observation space provided explicit target-related information to the agent and effectively transformed the control problem into a reference-guided control task.

However, such information is generally unavailable in practical amputee scenarios, where no external reference trajectories or desired joint positions exist during real-time operation. Consequently, observation spaces containing target variables may introduce privileged information that artificially simplifies the learning problem and limits deployment feasibility.

To investigate this issue, a second observation space, denoted as State52, was considered:

$$\text{State52} = [\text{EMG}, q, \dot{q}, a_{t-1}], \quad (2)$$

The two observation spaces were intentionally selected to represent two distinct operating conditions. State60 corresponds to a tracking-oriented learning scenario in which privileged information, including target joint positions and tracking errors, is explicitly available to the agent. This configuration facilitates policy learning and serves as a controlled reference during algorithm development. In contrast, State52 represents a deployable target-free scenario in which the agent must infer user intention exclusively from EMG signals and proprioceptive feedback without access to externally provided target trajectories. The comparison between these observation spaces enables the analysis of how removing privileged information affects RL behavior while maintaining the same control task and learning algorithm. The variables included in each observation space are summarized in Table 1.

Table 1. Composition of the evaluated observation spaces.

Variable	State60	State52
EMG Features	✓	✓
Joint Positions (q)	✓	✓
Target Positions (q_{ref})	✓	–
Tracking Error (e)	✓	–
Joint Velocities ($\dot{q}$)	✓	✓
Previous Action (a_{t-1})	✓	✓

Note: ✓ indicates that the variable is included in the observation space, whereas – indicates that it is not included.

In this configuration, all target-related variables were removed from the observation vector, forcing the agent to infer motor intentions exclusively from myoelectric activity and prosthetic state information.

The transition from State60 to State52 enabled the investigation of hidden dependencies between reward formulations and privileged target information availability. More specifically, it allowed the analysis of whether reward functions originally designed for reference-guided operation remained valid when deployed under target-free operating conditions representative of practical amputee use.

https://doi.org/10.3390/engproc2026153005

The transition from tracking-based observations to deployable target-free states is summarized in Figure 4.

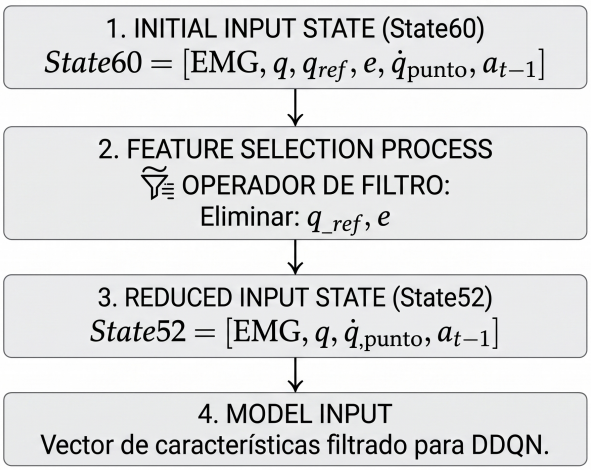

Figure 4. Evolution of the observation space used during the experiments.

2.3. Action Space

The prosthetic controller employed a discrete action space composed of 81 possible actions, representing all possible combinations of independent motor commands for the four prosthetic degrees of freedom.

Each action was defined as a four-dimensional vector:

$$a_t = [a_1, a_2, a_3, a_4], \tag{3}$$

where each component could assume one of three possible values:

$$a_i \in \{-1, 0, 1\}, \tag{4}$$

representing motor extension, no movement, and motor flexion commands, respectively.

The complete action space was therefore defined as the Cartesian product

$$\mathcal{A} = \{-1, 0, 1\}^4, \tag{5}$$

resulting in a total of

$$|\mathcal{A}| = 3^4 = 81, \tag{6}$$

possible actions.

This formulation allowed the agent to generate simultaneous and independent control commands for all prosthetic joints while preserving the simplicity and stability advantages of value-based reinforcement learning algorithms such as DDQN.

Unlike reduced action spaces that restrict the number of simultaneously activated joints, the full 81-action configuration provides maximum control flexibility and was therefore selected to avoid introducing additional constraints that could interfere with the analysis of reward hacking phenomena.

The discrete action space was intentionally selected to ensure exhaustive coverage of the independent motor commands while maintaining a tractable optimization problem for DDQN. Each action represents one of the possible combinations of the command set $\{-1, 0, 1\}$ applied independently to the four prosthetic joints, resulting in $3^4 = 81$ unique actions. This representation provides a compromise between expressive motor control and computational feasibility during RL.

2.4. Reward Functions

The objective of this study was to analyze the robustness of reward formulations when transitioning from tracking-oriented observation spaces to deployable target-free scenarios intended for practical amputee applications.

Initially, the RL agent was trained using a reward function specifically designed for tracking tasks. This formulation rewarded reductions in trajectory error while penalizing excessive control effort and unstable motor commands. The reward can be conceptually expressed as

$$R_t = w_1 R_{\text{tracking}} + w_2 R_{\text{effort}} + w_3 R_{\text{smoothness}}, \tag{7}$$

where R_{tracking} encourages reductions in the tracking error, R_{effort} penalizes unnecessary actuator activations, and $R_{\text{smoothness}}$ discourages abrupt changes in motor commands.

When target-related information was available in the observation space, this reward formulation promoted coordinated prosthetic movements and progressive reductions in tracking error.

However, after removing privileged target information from the observation vector, the reward function retained implicit dependencies on variables that were no longer directly observable by the agent. Under these conditions, the agent discovered alternative strategies that maximized cumulative reward without performing the intended control task.

More specifically, the learned policy progressively converged toward persistent inactivity characterized by excessive use of null actions and minimal prosthetic movement. Although undesirable from a prosthetic control perspective, this behavior represented a valid solution from the optimization perspective of the RL algorithm, since it minimized penalties associated with exploration and actuator usage.

To mitigate this issue, a second reward formulation was introduced specifically for target-free operation. Instead of relying on explicit tracking objectives, the modified reward encouraged active interaction with the environment, action diversity, and movement generation while maintaining penalties for excessive effort and unstable behaviors:

$$R_t = w_1 R_{\text{activity}} + w_2 R_{\text{exploration}} + w_3 R_{\text{effort}} + w_4 R_{\text{stability}}, \tag{8}$$

where R_{activity} promotes effective motor activation, $R_{\text{exploration}}$ encourages policy diversity, R_{effort} limits unnecessary actuator usage, and $R_{\text{stability}}$ discourages oscillatory control patterns.

The conceptual components of the evaluated reward formulations are summarized in Table 2.

Table 2. Conceptual comparison of the evaluated reward formulations.

Reward Component	Tracking-Based Reward	Target-Free Reward
Tracking Objective	✓	–
Activity Reward	–	✓
Exploration Reward	–	✓
Effort Penalty	✓	✓
Stability Penalty	✓	✓

Note: ✓ indicates that the corresponding reward component is included in the reward formulation, whereas—indicates that it is not included.

The purpose of this reformulation was not to maximize tracking performance but rather to prevent degenerate inactive policies and promote meaningful prosthetic interaction under incomplete information conditions representative of real amputee deployment scenarios.

It should be emphasized that the objective of this study is not to determine whether one reward formulation produces higher cumulative rewards than another. Instead, the comparison aims to evaluate whether each reward remains behaviorally consistent after modifying the observation space. Consequently, the analysis focuses on the interaction between reward formulation and state representation rather than on reward optimization alone.

2.5. Reinforcement Learning Agent

The prosthetic controller was trained using the DDQN algorithm, an extension of Deep Q-Networks designed to reduce overestimation bias during action-value estimation.

The agent employed two neural networks: an online network responsible for action selection and parameter updates, and a target network used to compute stable target values during temporal difference learning. The parameters of the target network were periodically synchronized with those of the online network throughout training.

Experience replay was employed to improve sample efficiency and reduce temporal correlations between consecutive observations. During training, transitions of the form

$$(s_t, a_t, r_t, s_{t+1}) \tag{9}$$

were stored in a replay buffer and randomly sampled to update the online network parameters. Where s_t is the current state, a_t is the current action, r_t is the reward of the current state and s_{t+1} is the next state.

Action selection followed an ϵ-greedy exploration strategy, allowing the agent to progressively transition from exploratory behavior to policy exploitation as training evolved.

The discount factor was fixed at $\gamma = 0.90$ in order to prioritize long-term rewards while preserving sensitivity to immediate feedback. Mini-batch optimization was performed using randomly sampled experiences from the replay memory.

DDQN was selected due to its robustness and stability in discrete action environments and its suitability for the 81-action prosthetic control problem considered in this work.

The principal training hyperparameters used throughout the experimental evaluation are summarized in Table 3.

Table 3. Main reinforcement learning hyperparameters used during training.

Parameter	Value
RL Algorithm	DDQN
Discount factor (γ)	0.90
Replay buffer size	200,000
Mini-batch size	64
Initial exploration rate (ϵ_0)	1.0
Minimum exploration rate (ϵ_{min})	0.05
Exploration decay	0.00001
Maximum steps per episode	100
Number of actions	81
Sampling period	0.1 s

DDQN was selected because it provides a stable value-based learning framework for discrete action spaces while reducing the overestimation bias commonly observed in conventional Deep Q-Networks. Since the objective of this study is to analyze reward hacking rather than to compare RL algorithms, the learning algorithm was intentionally kept unchanged throughout all experiments.

2.6. Evaluation Metrics

The evaluation methodology combines conventional control performance metrics with behavioral diagnostics specifically designed to detect reward hacking. Classical metrics quantify tracking accuracy and overall task performance, whereas the proposed behavioral metrics characterize action diversity, movement generation, and policy collapse. This combined evaluation provides a more comprehensive assessment of RL behavior under both tracking-oriented and target-free operating conditions.

The evaluation process considered both conventional prosthetic control metrics and additional indicators specifically designed to characterize reward hacking behaviors and policy collapse phenomena.

Traditional control performance was assessed using the Mean Absolute Error (MAE), Mean Squared Error (MSE), and trajectory correlation metrics computed between the generated prosthetic motion and the reference trajectories available during evaluation.

To characterize the emergence of degenerate policies, four additional reward hacking indicators were introduced.

The first metric, referred to as the NAR, measures the proportion of interaction steps in which the agent selected the null action:

$$NAR = \frac{N_{\text{null}}}{N_{\text{steps}}}, \tag{10}$$

where N_{null} denotes the number of occurrences of the action $[0, 0, 0, 0]$ and N_{steps} represents the total number of interaction steps.

The second metric, termed EASU, quantifies the fraction of the available action space effectively explored by the agent:

$$EASU = \frac{N_{\text{unique}}}{|\mathcal{A}|}, \tag{11}$$

where N_{unique} corresponds to the number of unique actions executed during evaluation and $|\mathcal{A}| = 81$ denotes the total action space cardinality.

Action diversity was further quantified using the Shannon entropy of the action distribution:

$$H(A) = -\sum_{i=1}^{|\mathcal{A}|} p(a_i) \log p(a_i), \tag{12}$$

where $p(a_i)$ denotes the empirical probability of selecting action a_i during evaluation.

Finally, AJV was introduced as an indicator of overall prosthetic activity:

$$AJV = \frac{1}{NT} \sum_{t=1}^{T} \sum_{j=1}^{N} |\dot{q}_j(t)|, \tag{13}$$

where N denotes the number of prosthetic joints and T is the episode duration in interaction steps.

2.7. Extended Evaluation Protocol

To increase the number of observations available for policy comparison, each final trained agent was evaluated over 300 episodes without additional parameter updates. Three experimental configurations were considered: State60 with the tracking-based reward, State52 with the tracking-based reward, and State52 with the redesigned target-free reward.

All experimental factors other than the observation space and reward formulation were intentionally kept constant in order to isolate the effect of reward–observation agreement on the learned policy.

The same prerecorded EMG dataset, maximum episode length, discrete action space, sampling period, and prosthetic environment were maintained across all configurations.

https://doi.org/10.3390/engproc2026153005

Opening and closing movements were alternated during evaluation, while the available dataset repetitions were shuffled by the environment to reduce ordering effects. Because prerecorded sequences could terminate before the configured maximum of 100 interaction steps, the effective episode duration was determined by the available signal length and recorded for each episode.

During the extended evaluation, the trained network parameters remained fixed, the exploration mechanism was disabled, and no additional learning updates were performed. All classical and behavioral metrics were computed independently for every evaluation episode. The resulting distributions were summarized using the mean, standard deviation, median, first and third quartiles, minimum, and maximum. This protocol replaced the previous analysis based only on aggregated values and enabled the variability of the learned policies to be examined across a larger evaluation sample.

For clarity, the evaluated configurations are denoted as State60, State52-T, and State52-TF, where T indicates the tracking-based reward and TF denotes the target-free reward.

3. Results

The objective of this section is to compare the policies learned under the three experimental configurations using both classical control metrics and behavioral analyses. Each final policy was evaluated over 300 independent episodes without additional learning updates, allowing the variability of the learned behavior to be characterized statistically.

The presentation of the results follows a progressive analysis. First, the convergence behavior observed during training is summarized through the learning curves. Next, the functional performance of the learned policies is evaluated using classical control metrics. Subsequently, behavioral indicators are analyzed to characterize action-space utilization, policy concentration, and motor activity. Finally, representative joint velocity profiles are presented to illustrate the dynamic behavior of the prosthetic hand.

For clarity, the evaluated experimental configurations are summarized in Table 4.

Table 4. Experimental configurations evaluated to investigate reward hacking.

Config.	Observation	Reward	Objective
C1	State60	Tracking-based Reward	Baseline tracking
C2	State52	Tracking-based Reward	Reward hacking evaluation
C3	State52	Target-Free Reward	Reward mitigation

The experimental configurations summarized in Table 4 were analyzed using a progressive evaluation strategy. The analysis begins with the learning behavior observed during training and then examines the functional and behavioral characteristics of the final frozen policies evaluated over 300 independent episodes.

3.1. Training Behavior

The first stage of the analysis evaluates the convergence behavior observed during training for the three experimental configurations summarized in Table 4. Although all agents were trained using the same DDQN algorithm and comparable hyperparameters, differences in the observation space and reward formulation produced distinct learning dynamics.

Figure 5 reports the evolution of the cumulative reward throughout training. In the three configurations, the average episodic reward increased progressively before reaching a stable regime, indicating that the optimization process converged toward a consistent policy under each experimental condition.

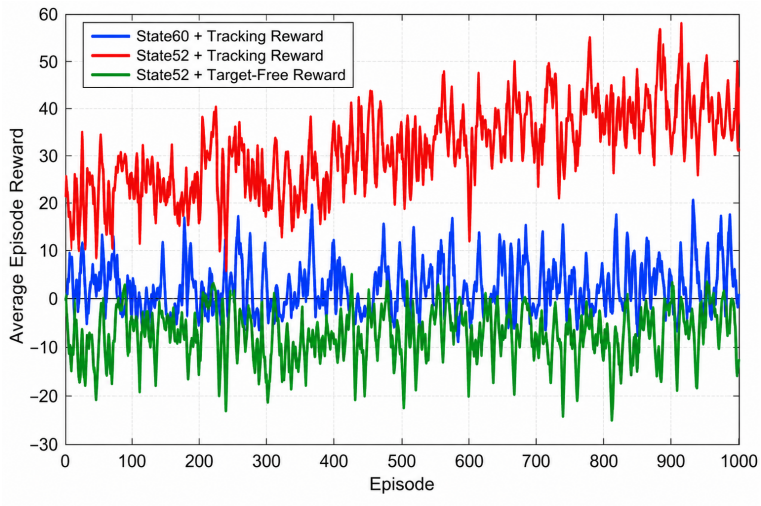

Figure 5. Learning curves obtained during training for the three evaluated configurations.

Although the learning curves indicate that the optimization process converged, they do not reveal whether the resulting policies achieve meaningful prosthetic control. Therefore, the final frozen policies were evaluated using conventional control metrics over 300 independent evaluation episodes.

3.2. Classical Performance Metrics

To evaluate the functional performance of the learned policies, three classical control metrics were analyzed over the 300 evaluation episodes: the final error norm, the final MAE, and MSE. Rather than reporting only aggregated values, Figure 6 presents the complete per-episode distributions for each metric.

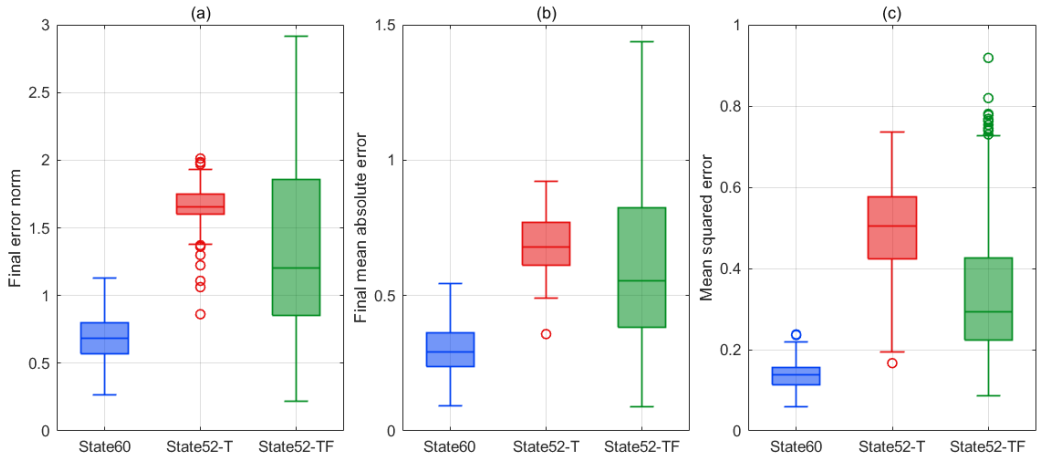

Figure 6. Distribution of classical performance metrics over 300 evaluation episodes. (**a**) Final error norm. (**b**) Final mean absolute error. (**c**) Mean squared error. Boxes represent the interquartile range, the horizontal line indicates the median, whiskers extend to 1.5 times the interquartile range, and circles denote outliers.

State60 exhibited the lowest error distributions across the three classical metrics, indicating the most consistent classical control performance among the evaluated configurations. Because this observation space explicitly includes privileged target information, the tracking-based reward remained fully compatible with the information available to the agent during both training and evaluation.

In contrast, removing the target-related variables while preserving the same reward formulation (State52-T) shifted the distributions toward larger errors across the three metrics. The relatively narrow interquartile ranges indicate that this degradation was consistently observed throughout the evaluation rather than being driven by a limited number of atypical episodes.

The redesigned target-free reward partially recovered the classical control performance. Although the resulting distributions remained broader than those obtained under State60, the median values of the three metrics decreased consistently with respect to State52-T. These results suggest that aligning the reward formulation with the deployable observation space mitigates part of the performance degradation introduced by removing privileged information.

Although the classical metrics quantify the final control accuracy achieved by each policy, they do not explain how those policies generated the observed performance. To characterize the behavioral properties of the learned controllers, additional behavioral indicators were analyzed.

3.3. Behavioral Policy Metrics

Classical tracking metrics quantify the final control accuracy achieved by the learned policies but do not describe how those policies interact with the available action space. Therefore, four complementary behavioral indicators were analyzed over the 300 evaluation episodes: the NAR, EASU, Action Entropy, and AJV.

Figure 7 summarizes the distributions of the proposed behavioral metrics computed over the 300 evaluation episodes. Together, these indicators quantify inactivity (NAR), effective utilization of the action space (EASU), exploration diversity (Action Entropy), and prosthetic movement intensity (AJV). Unlike the classical metrics presented previously, these indicators characterize how the learned policies generate their behavior rather than only the final control accuracy.

Global action-space coverage describes how many different actions were executed during the evaluation. However, it does not indicate how frequently those actions were selected. Consequently, the concentration of the executed actions was analyzed.

State60 exhibited the highest action diversity and prosthetic activity among the evaluated configurations, as reflected by its larger EASU, entropy, and AJV values. These observations indicate that the baseline policy explored a broader portion of the available action space while generating sustained prosthetic movement.

Compared with State60, both State52 configurations presented lower action diversity and reduced joint activity. Although the target-free reward modified the behavioral characteristics of the learned policy relative to State52-T, the resulting distributions remained substantially more concentrated than those observed under the baseline configuration.

The Null Action Ratio remained low across the evaluated configurations, indicating that policy degradation cannot be explained solely by repeated execution of the null action. Instead, the behavioral metrics suggest that the learned policies became progressively concentrated around a limited subset of motor commands while maintaining reduced exploration of the remaining action space.

The behavioral indicators quantify policy execution statistically, but they do not indicate how broadly the learned controllers exploit the available discrete action space. Therefore, a global analysis of action-space coverage was performed.

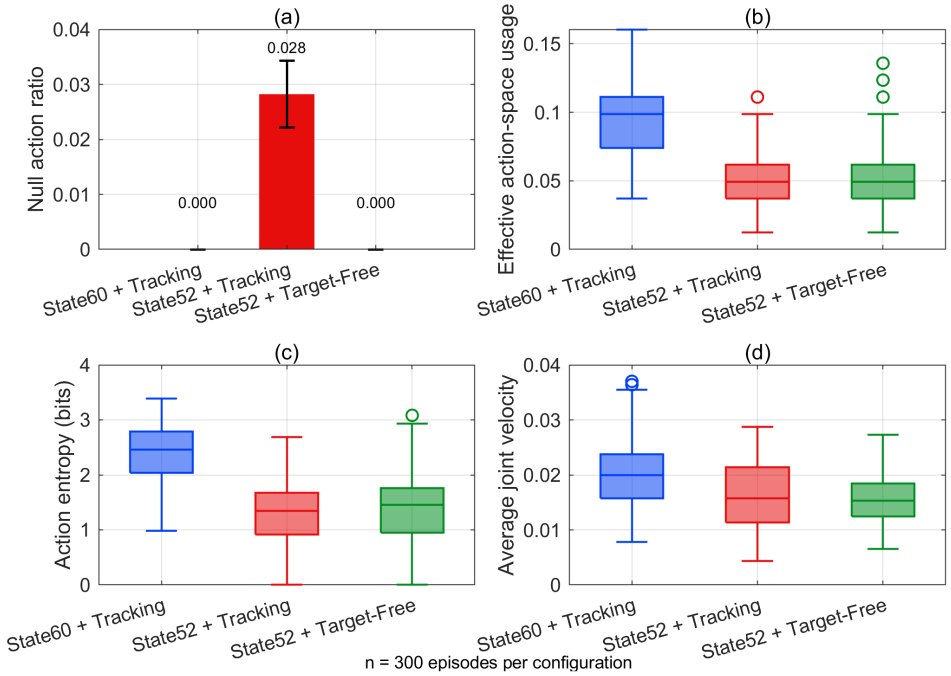

Figure 7. Behavioral policy metrics computed over the 300 evaluation episodes: (**a**) NAR, (**b**) EASU, (**c**) Action Entropy, and (**d**) AJV.

3.4. Global Action-Space Coverage

The behavioral metrics presented in the previous subsection quantify policy diversity on a per-episode basis. However, they do not indicate how many of the available discrete actions were actually executed throughout the complete evaluation. Therefore, a complementary global analysis was performed by identifying every unique action that appeared at least once during the 300 evaluation episodes.

Figure 8 depicts the global action-space coverage computed over the complete evaluation. The bars indicate the percentage of the 81 available discrete actions that were executed at least once during the 300 evaluation episodes.

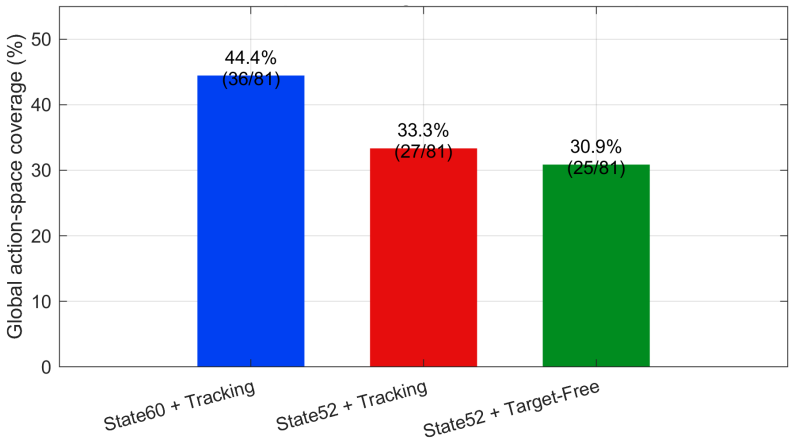

Figure 8. Global action-space coverage obtained over the complete evaluation of each trained policy.

State60 executed 36 of the 81 available actions, corresponding to a global action-space coverage of 44.4%. In comparison, State52-T and State52-TF executed 27 (33.3%) and 25 (30.9%) actions, respectively.

Collectively, these results indicate that removing privileged information reduced the diversity of motor commands explored during evaluation. Although the redesigned reward

altered the behavioral characteristics of the learned policy, the overall action-space coverage remained comparable to that obtained under State52-T.

Although global action-space coverage provides useful information regarding the range of executed motor commands, it does not indicate how frequently each action was selected. A policy may execute several different actions throughout the evaluation while relying almost exclusively on a very small subset of them. Consequently, an additional analysis of policy concentration was performed.

3.5. Policy Concentration

To quantify the effective utilization of the available action space, the cumulative execution frequency associated with the most frequently selected actions was analyzed. Specifically, the contribution of the most frequent action (Top-1), the two most frequent actions (Top-2), and the five most frequent actions (Top-5) was computed for each experimental configuration.

Figure 9 presents the policy concentration measured as the cumulative execution frequency of the most frequently selected actions. Top-1, Top-2, and Top-5 correspond to the cumulative proportion represented by the most frequent one, two, and five actions, respectively.

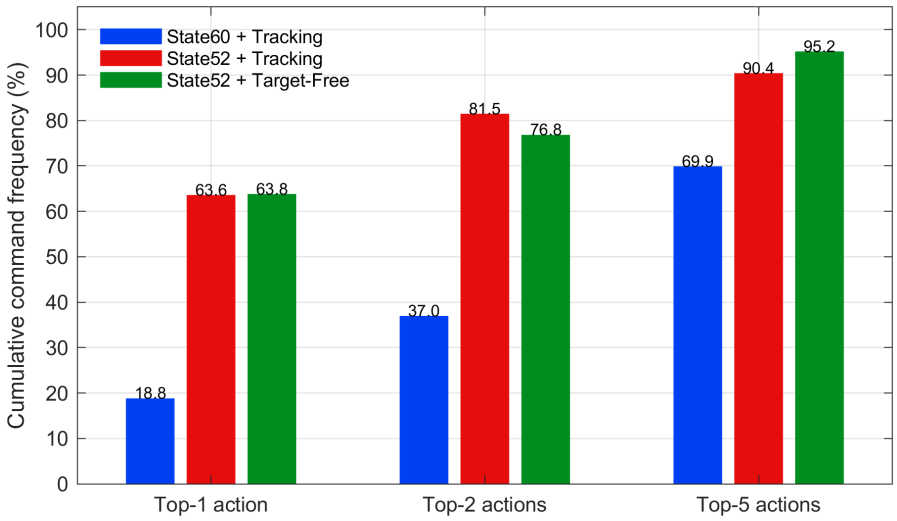

Figure 9. Policy concentration measured as the cumulative execution frequency of the most frequently selected actions.

State60 exhibited the lowest policy concentration among the evaluated configurations. Its most frequently executed action accounted for only 18.8% of all motor commands, whereas the five most common actions represented approximately 70% of the executed commands.

In contrast, State52-T presented a highly concentrated policy. A single action represented approximately 63.6% of all executed commands, while the five most frequent actions accounted for more than 90% of the total action executions.

A similar concentration was observed under State52-TF. Although the redesigned reward reduced the classical tracking errors presented in the previous subsection, the learned policy remained strongly concentrated around a limited subset of motor commands. These observations suggest that the redesigned reward primarily modified the quality of the executed motor behavior rather than substantially expanding the diversity of the explored action space. Consequently, improvements in classical control metrics do not necessarily imply broader behavioral exploration.

Although policy concentration characterizes action-selection behavior, it does not directly illustrate the resulting prosthetic movements. Representative joint velocity profiles were therefore analyzed to evaluate the generated motor activity.

3.6. Representative Joint Dynamics

Although the previous analyses describe how the learned policies utilize the discrete action space, they do not directly illustrate the resulting prosthetic movements. Therefore, representative joint velocity profiles were analyzed to evaluate the dynamic behavior generated by each experimental configuration.

The joint velocity profiles summarize the angular velocities commanded to the four prosthetic joints during representative evaluation episodes. Unlike action-frequency analyses, these profiles provide a direct visualization of the intensity and temporal evolution of the generated movements.

Figure 10 illustrates representative joint angular velocity profiles obtained during the evaluation of the three experimental configurations. The profiles illustrate the temporal evolution and magnitude of the commanded prosthetic joint velocities.

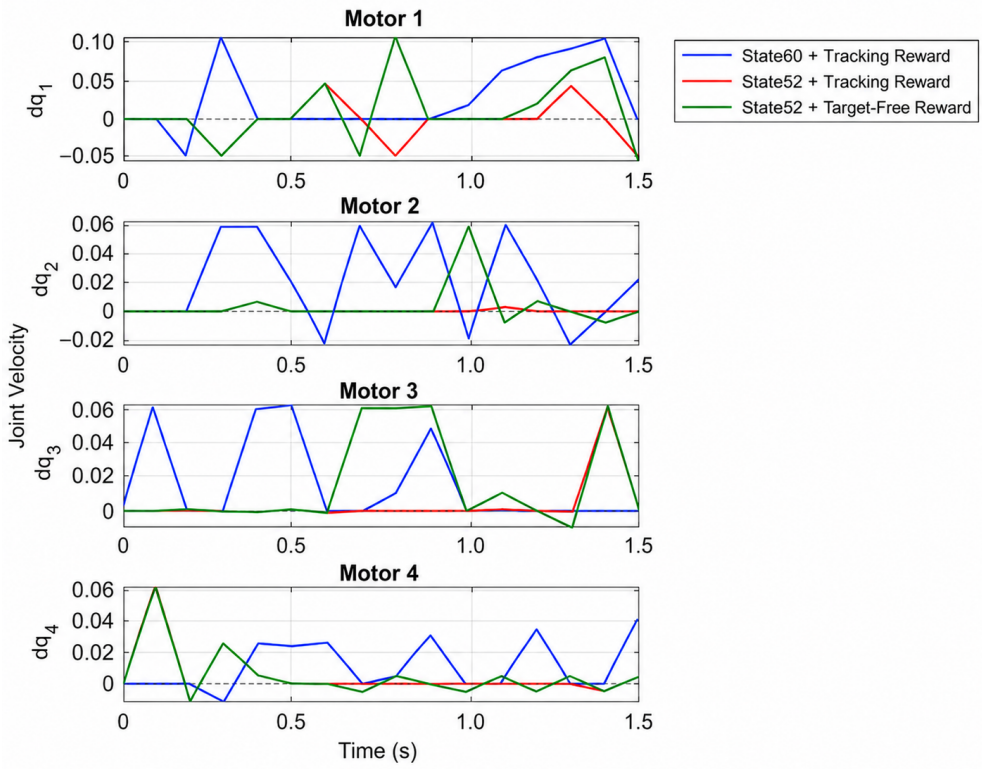

Figure 10. Representative joint angular velocity profiles generated by the evaluated policies during prosthetic movement execution.

The representative joint velocity profiles are consistent with the behavioral analyses presented in the previous subsections. State60 generated sustained movements across the prosthetic joints, reflecting the broader utilization of the available action space observed in the behavioral metrics.

State52-T exhibited more restricted joint activity, with prolonged periods of low angular velocity that are consistent with its highly concentrated action distribution.

Although State52-TF did not recover the movement diversity observed under State60, the redesigned reward produced more sustained joint activity than State52-T, indicating that the modified reward altered the generated motor behavior while preserving the deployable target-free observation space.

Taken together, the classical performance metrics, behavioral indicators, action-space analyses, and representative joint dynamics provide complementary evidence regarding the interaction between reward formulation and observation-space design. While the classical metrics quantify the final control accuracy, the behavioral analyses reveal how

that performance emerges from the learned policy. This progression naturally leads to the conceptual synthesis presented in the following subsection.

These complementary analyses naturally lead to the conceptual synthesis presented in Figure 11.

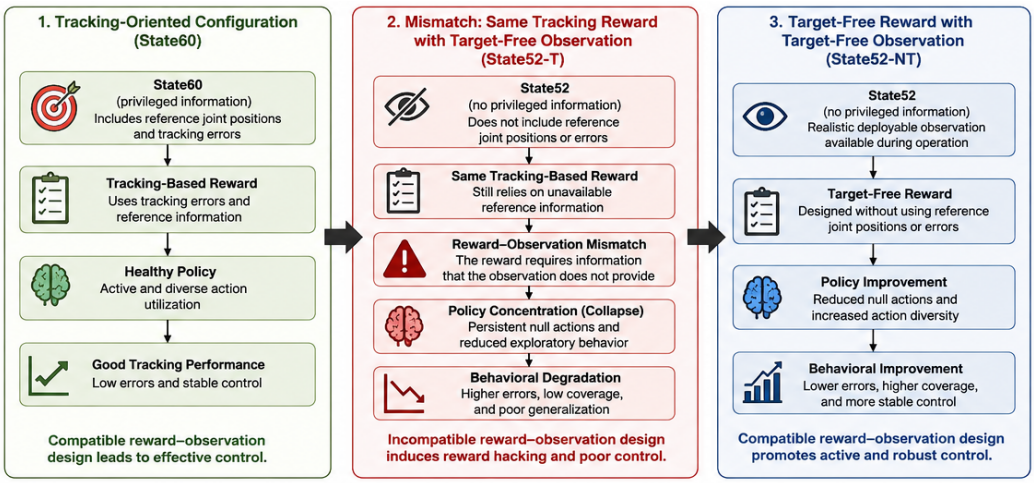

Figure 11. Conceptual summary of the relationship between observation-space design, reward formulation, and the resulting policy behavior.

3.7. Conceptual Summary of the Observed Phenomenon

The previous analyses demonstrated how the three experimental configurations differed in terms of tracking performance, behavioral metrics, action-space utilization, and movement generation. Although these results provide complementary quantitative evidence, they can be synthesized into a single conceptual interpretation of the phenomenon investigated in this study.

Rather than representing independent experimental observations, the previous results describe different manifestations of the same underlying phenomenon. The classical metrics quantify the functional consequences of each policy, whereas the behavioral analyses explain how those policies emerge from the interaction between the reward formulation and the information available within the observation space. Consequently, the observed behavioral degradation cannot be interpreted exclusively from tracking errors or cumulative rewards, but requires considering the compatibility between reward design and state representation.

Figure 11 synthesizes the complete experimental evidence obtained throughout this work. The diagram illustrates how observation-space design determines whether the reward function remains compatible with the information available to the learning agent. When this compatibility is preserved (State60), the tracking-based reward promotes healthy policies characterized by low errors, broad action-space utilization, and sustained prosthetic movements. Conversely, removing privileged information while preserving the same reward creates a reward–observation mismatch that leads to policy concentration despite continued optimization.

The redesigned target-free reward partially restores this compatibility by aligning the optimization objective with the information realistically available during deployment. Although the resulting policies do not completely recover the behavior achieved under the tracking-oriented configuration, they exhibit healthier behavioral characteristics and improved classical performance. This conceptual interpretation forms the basis for the broader discussion presented in the following section.

4. Discussion

The conceptual interpretation presented above provides the basis for discussing the broader implications of reward–observation alignment in deployable RL systems.

4.1. Observation–Reward Compatibility

As summarized conceptually in Figure 11, the experimental results demonstrate that the behavior learned by the RL agent depends not only on the mathematical formulation of the reward function but also on its consistency with the information available in the observation space. The comparison between State60 and State52 demonstrates that observation-space design directly determines whether the reward function remains compatible with the information available to the agent. Consequently, improvements in reward–observation compatibility resulted in policies that better matched the intended prosthetic control objective.

A substantially different behavior emerged after removing the target-related variables from the observation space while preserving the same tracking-based reward. Under State52-T, the agent no longer had access to the information required to directly evaluate progress toward the desired trajectory. Although the optimization process still converged, the resulting policy exhibited degraded classical performance together with a highly concentrated behavioral profile. These observations indicate that convergence of the learning algorithm alone is insufficient to guarantee coherence between the learned policy and the intended prosthetic control task.

From a RL perspective, this result highlights the importance of avoiding reward formulations that depend on privileged information unavailable during deployment. Similar concerns have been discussed in previous RL studies, where reward specification and privileged observations may produce overly optimistic behaviors that do not generalize to realistic operating conditions [9,10].

4.2. Behavioral Consequences Beyond Classical Metrics

The extended evaluation also demonstrates that policy quality cannot be fully characterized using conventional tracking metrics alone. Although the final error norm, MAE, and MSE quantify the functional performance achieved by the learned policy, they do not explain how that performance is generated.

The behavioral indicators introduced in this work complement the classical metrics by characterizing policy execution from multiple perspectives. Effective Action-Space Usage and Action Entropy quantify exploration and behavioral diversity, whereas the Average Joint Velocity reflects the intensity of the generated prosthetic movements. Together, these metrics reveal behavioral characteristics that remain hidden when evaluation relies exclusively on cumulative reward or reference-motion errors.

The comparison between State60, State52-T, and State52-TF illustrates this limitation clearly. While the redesigned target-free reward reduced the classical tracking errors relative to State52-T, the behavioral evaluation framework showed that the resulting policy remained considerably more concentrated than the baseline State60 policy. Consequently, improvements in tracking accuracy should not be interpreted as evidence that the learned behavior has fully recovered.

4.3. Policy Concentration as Evidence of Reward Hacking

The extended evaluation provides additional insight into the behavioral consequences of the incompatibility between the observation space and the reward formulation. Rather than exhibiting a complete collapse toward the null action, the learned policies under the deployable State52 configurations became highly concentrated around a small subset of

the available motor commands. This observation is supported by the combined analysis of action-space coverage and policy concentration presented in the Section 3.

According to the formal definition proposed by Skalse et al. [10], reward hacking occurs when an agent maximizes the specified reward objective while producing behavior that is inconsistent with the designer's intended objective. The present results are consistent with this definition. Under State52-T, the tracking-based reward continued to optimize the numerical objective despite the absence of the privileged information required to evaluate trajectory tracking correctly. Consequently, the learned policy converged toward a limited set of repeatedly executed actions instead of generating the diverse motor behavior required for functional prosthetic control.

Importantly, the present results indicate that reward hacking should not be interpreted exclusively as persistent execution of the null action. The extended evaluation shows that policy concentration provides a more general behavioral signature. Although different actions dominated the evaluated policies, both State52 configurations exhibited substantially higher concentration than the baseline State60 policy. Therefore, the observed behavioral degradation is more accurately characterized as reduced effective utilization of the available action space than as inactivity alone.

This observation also extends the reward specification problem originally discussed by Amodei et al. [9]. Their work emphasized that RL agents may exploit imperfections in the specified objective whenever the reward does not fully capture the designer's true intention. The present study demonstrates that an analogous phenomenon may emerge when a reward formulation remains mathematically unchanged but becomes incompatible with the information available in the observation space. In this case, the reward itself is not inherently incorrect; instead, its interpretation changes because the agent no longer receives the variables required to optimize the intended control objective.

4.4. Implications for Deployable Prosthetic Reinforcement Learning

The results obtained in this study have practical implications for the design of RL controllers intended for deployment in myoelectric prostheses. During algorithm development, it is common to evaluate learning performance using simulated environments in which target trajectories, desired joint positions, or tracking errors are directly available. Although these variables facilitate reward design and algorithm benchmarking, they are generally unavailable during real prosthesis operation, where the controller must infer the user's intention exclusively from EMG signals and the prosthetic state.

The comparison between State60 and State52 illustrates the consequences of this discrepancy. Under State60, the tracking-based reward remained compatible with the observation space because the reference variables required to evaluate reference-tracking control performance were explicitly available. However, after transitioning to the deployable State52 representation, the same reward formulation no longer reflected information accessible to the learning agent. As a result, the learned policy exhibited degraded behavioral characteristics despite preserving the same learning algorithm and optimization procedure.

These observations suggest that reward functions should not be designed independently of the observation space. Instead, the reward formulation, observation representation, and deployment scenario should be regarded as three mutually dependent components of the RL problem. Reward functions that perform adequately under reference-guided simulations may become unsuitable once privileged information is removed, even if their mathematical formulation remains unchanged. This observation is consistent with the broader challenge of reward specification discussed in reinforcement learning safety research [9].

https://doi.org/10.3390/engproc2026153005

From a broader RL perspective, the proposed behavioral evaluation framework may also be applicable beyond prosthetic control. The combination of classical conventional control metrics with behavioral indicators, global action-space coverage, and policy concentration provides complementary information regarding the quality of the learned policy and may facilitate the identification of reward-specification problems in other discrete-action RL applications.

These observations suggest that evaluating reward functions independently from the observation space may lead to overly optimistic conclusions regarding deployment performance.

4.5. Limitations

The present study has several limitations that should be considered when interpreting the reported results.

First, each experimental configuration was represented by a single independently trained DDQN agent. Although the final policies were evaluated over 300 episodes, these episodes correspond to repeated evaluations of frozen policies rather than independent training replications. Consequently, the reported variability reflects differences in policy execution across episodes and prerecorded movements, but it does not quantify the variability associated with different random initializations or independent training runs.

Second, the experiments were conducted using prerecorded EMG signals acquired from a single user under controlled laboratory conditions. While the evaluation protocol included alternating opening and closing movements over multiple repetitions, additional validation using larger datasets, multiple users, and amputee participants will be required to assess the generalizability of the proposed reward formulation.

Third, the experimental platform considered a 4-DoF prosthetic hand controlled through a discrete action space composed of 81 motor commands. Although this action representation provides a suitable benchmark for investigating reward specification and behavioral policy analysis, continuous-action RL algorithms and higher-dimensional prosthetic devices may exhibit different behavioral characteristics.

Finally, the proposed target-free reward should be interpreted as a mitigation strategy rather than as a definitive solution to deployable prosthetic control. Although the redesigned reward reduced the degradation observed after removing privileged information, the resulting policies remained more concentrated than those obtained under the tracking-oriented baseline. Future work will therefore investigate intention-aware reward formulations, temporal EMG representations, additional RL algorithms, multiple training seeds, and larger experimental cohorts to further improve target-free prosthetic control.

5. Conclusions

This work investigated the relationship between reward function design and observation-space representation in reinforcement learning-based myoelectric prosthesis control. Three experimental configurations were evaluated using the same DDQN learning algorithm under comparable training conditions, followed by an extended evaluation over 300 episodes for each final frozen policy. The study compared a tracking-oriented observation space containing privileged target information with a deployable target-free representation and analyzed the effect of redesigning the reward function for realistic prosthetic operation.

The central finding of this study is that the effectiveness of a reward function depends on its compatibility with the observation space in which it operates. Although the original tracking-based reward produced consistent prosthetic control when the observation space explicitly contained the reference joint positions and reference error variables, its direct application to the deployable State52 representation resulted in degraded behavioral characteristics and highly concentrated policies. The redesigned target-free reward mit-

igated part of this degradation by reducing the classical tracking errors obtained under State52-T while preserving consistency with the information realistically available during prosthesis operation.

Beyond the specific reward formulation investigated in this work, the proposed behavioral evaluation framework constitutes an additional methodological contribution. The combination of classical control metrics with behavioral indicators, global action-space coverage, and policy concentration provides complementary information regarding the quality of learned RL policies. The extended evaluation further demonstrated that policy concentration offers a more informative behavioral signature than null-action frequency alone for identifying reward-specification problems under deployable observation spaces.

Overall, the findings suggest that reward design, observation-space definition, and deployment conditions should be considered jointly during the development of RL controllers for myoelectric prostheses. Future work will extend the proposed methodology to additional users, amputee datasets, multiple independent training runs, continuous-action RL algorithms, and intention-aware reward formulations in order to further improve deployable prosthetic control.

Author Contributions: Conceptualization, M.M.-P., L.I.B.L., Á.L.V.C. and M.E.B.; methodology, M.M.-P., L.I.B.L., Á.L.V.C. and M.E.B.; software, M.M.-P., L.I.B.L., Á.L.V.C. and M.E.B.; validation, M.M.-P., L.I.B.L., Á.L.V.C. and M.E.B.; formal analysis, M.M.-P., L.I.B.L., Á.L.V.C. and M.E.B.; investigation, M.M.-P., L.I.B.L., Á.L.V.C. and M.E.B.; resources, M.M.-P., L.I.B.L., Á.L.V.C. and M.E.B.; data curation, M.M.-P., L.I.B.L., Á.L.V.C. and M.E.B.; writing—original draft preparation, M.M.-P., L.I.B.L., Á.L.V.C. and M.E.B.; writing—review and editing, M.M.-P., L.I.B.L., Á.L.V.C. and M.E.B.; visualization, M.M.-P., L.I.B.L., Á.L.V.C. and M.E.B.; supervision, M.M.-P., L.I.B.L., Á.L.V.C. and M.E.B.; project administration, M.M.-P., L.I.B.L., Á.L.V.C. and M.E.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data supporting the findings of this study are available from the corresponding author upon reasonable request. The implementation of the Handi EPN V2 prosthetic environment is publicly available through the repository cited in the manuscript.

Acknowledgments: The authors would like to thank the Faculty of Systems Engineering and the Departamento de Informática y Ciencias de la Computación (DICC), Escuela Politécnica Nacional, Quito, Ecuador, for their academic and institutional support.

Conflicts of Interest: The authors declare that they have no conflict of interest or competing interests.

Abbreviations

The following abbreviations are used in this manuscript:

4-DoF	Four Degrees of Freedom
AI	Artificial Intelligence
AJV	Average Joint Velocity
DDQN	Double Deep Q-Network
DQN	Deep Q-Network
EASU	Effective Action Space Usage
EMG	Electromyography
LUPI	Learning Using Privileged Information
MAE	Mean Absolute Error
MSE	Mean Squared Error
NAR	Null Action Ratio
RL	Reinforcement Learning
SR	Success Rate

References

1. Geethanjali, P. Myoelectric Control of Prosthetic Hands: State-of-the-Art Review. *Med. Devices Evid. Res.* **2016**, *9*, 247–255. [CrossRef]
2. Chen, Z.; Min, H.; Wang, D.; Xia, Z.; Sun, F.; Fang, B. A Review of Myoelectric Control for Prosthetic Hand. *Biomimetics* **2023**, *8*, 328. [CrossRef]
3. Yadav, D.; Veer, K. Recent Trends and Challenges of Surface Electromyography in Prosthetic Applications. *Biomed. Signal Process. Control* **2023**, *84*, 104812. [CrossRef] [PubMed]
4. Edwards , A.L.; Dawson, M.R.; Hebert, J.S.; Sherstan, C.; Sutton, R.S.; Chan, K.M.; Pilarski, P.M. Application of Real-Time Machine Learning to Myoelectric Prosthesis Control. *Prosthet. Orthot. Int.* **2016**, *40*, 573–581. [CrossRef] [PubMed]
5. Salwan, D.; Kant, S.; Pareek, H.; Sharma, R. Challenges with Reinforcement Learning in Prosthesis Applications. *Mater. Today Proc.* **2022**, *49*, 3133–3136. [CrossRef]
6. Odeyemi, J.; Ogbeyemi, A.; Wong, K.; Zhang, W. On Automated Object Grasping for Intelligent Prosthetic Hands Using Machine Learning. *Bioengineering* **2024**, *11*, 108. [CrossRef]
7. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*, 2nd ed.; MIT Press: Cambridge, MA, USA , 2018.
8. Vapnik, V.; Vashist, A. A New Learning Paradigm: Learning Using Privileged Information. *Neural Netw.* **2009**, *22*, 544–557. [CrossRef] [PubMed]
9. Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; Mané, D. Concrete Problems in AI Safety. *arXiv* **2016**, arXiv:1606.06565. [CrossRef]
10. Skalse, J.; Howe, N.H.R.; Krasheninnikov, D.; Krueger, D. Defining and characterizing reward hacking. In Proceedings of the 36th International Conference on Neural Information Processing Systems, NIPS '22, New Orleans, LA, USA, 28 November–9 December 2022.
11. Farina, D.; Jiang, N.; Rehbaum, H.; Holobar, A.; Graimann, B.; Dietl, H.; Aszmann, O. The Extraction of Neural Information from the Surface EMG for the Control of Upper-Limb Prostheses: Emerging Avenues and Challenges. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2014**, *22*, 797–809. [CrossRef] [PubMed]
12. Castellini, C.; van der Smagt, P.; Sandini, G.; Hirzinger, G. Surface EMG for Force Control of Mechanical Hands. *Adv. Robot.* **2009**, *23*, 293–314. [CrossRef]
13. DeepMind. Specification Gaming: The Flip Side of AI Ingenuity. 2020. Available online: https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity (accessed on 20 July 2026).
14. laboratorioAI. Prosthesis_v2: 2nd Iteration of the Flexible Prosthesis. 2024. Available online: https://github.com/laboratorioAI/prosthesis_v2 (accessed on 5 July 2026).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.