Reinforcement Learning for Pokémon Showdown

Jike Zhang, ajhz632@aucklanduni.ac.nz

1 Introduction

This report investigates how world design—specifically the definitions of *state*, *action*, and *reward*—influences the learning behavior of a reinforcement learning (RL) agent in the **Pokémon Showdown** environment. It focuses on how these design choices affect learning stability, effectiveness, and alignment with the intended objective.

The report first outlines the world design, including state representation, action constraints, and reward formulation. It then defines quantitative metrics—eg: *win rate*, *KO difference*, and *potential change* ($\Delta\Phi$)—to evaluate learning stability and performance. By comparing different world definitions, the study identifies trade-offs between *learnability* and *reward alignment*, and derives insights for constructing stable and interpretable RL environments.

Before discussing the design details, it is useful to recall the key challenges of reinforcement learning highlighted in prior work [4, 2, 1]—namely *partial observability*, *delayed rewards*, *large or constrained action spaces*, *reward misalignment*, and *stochasticity*. These difficulties are amplified in the **Pokémon Showdown** environment, where hidden opponent states, delayed outcomes, and random battle effects cause unstable learning.

These challenges are magnified in the **Pokémon Showdown** environment. ***Partial observability:*** hidden opponent Pokémon make full state information unavailable. ***Delayed rewards:*** the win signal appears only at the end of each battle, creating long credit-assignment chains. ***Constrained action space:*** many moves are invalid or context-dependent, increasing exploration complexity. ***Stochasticity:*** variable damage, accuracy, and status effects introduce high variance in outcomes. Together, these factors make Showdown a complex yet informative testbed for studying world and reward design in RL.

2 World Definition & Learning Framework

2.1 State Representation

Each battle turn is encoded as a **40-d** continuous vector capturing key decision factors (Table 1). Feature design follows three principles: *interpretability*, *learnability*, and *strategic relevance*. All variables are observable or estimable and their changes align with the reward signal. HP and KO count express survival, type and weather give context, speed captures initiative, and PP balances offense and resource use. All features are normalized to $[0, 1]$ or $[-1, 1]$ for stable scaling and learnability; integer type IDs reduce one-hot sparsity, and the summary variable Φ compresses correlated features into a compact representation.

Table 1: State vector (40-d): feature groups, design rationale, and role for learning.

Feature Group	Dim.	Description & Design Rationale	Role for Learning
HP	12	HP ratios for our and opponent teams (6+6). Core and stable signal of survival and pressure, reflecting resource and team health.	Key survival/offense cue; basis for potential (Φ) and reward evaluation.
Type	12	Integer IDs of Pokémon types for both sides. Types define offensive/defensive potential. Integer IDs reduce sparsity vs. one-hot.	Provides prior knowledge of type matchups; improves understanding of move effects.
Weather / Field	10	One-hot of current weather and terrain that globally alter damage and speed.	Enables perception of environment and adapts tactics under different conditions.
PP (avg.)	1	Average remaining PP, representing resource consumption and persistence.	Balances offense and resource saving.
Speed advantage	1	Whether we move first this turn (1=faster).	Represents initiative; affects risk and action choice.
KO count	2	Numbers of fainted Pokémon on each side; describe remaining team strength.	Reflects game phase and momentum; aids strategy adjustment.
Potential (Φ)	1	Aggregated score combining HP, type, speed, and weather, inspired by chess-like evaluation;	Provides dense directional feedback; compresses correlated features; links state to $\Delta\Phi$.
Turn ratio	1	Current turn index normalized by max turns, distinguishing early vs. late phase.	Allows adaptation of pacing across phases.

State redundancy and compactness. Designing a useful state space requires balancing *informational sufficiency*

and *computational compactness*. An overly rich state risks the curse of dimensionality, while an oversimplified one loses key decision cues. Although raw HP vectors theoretically describe the entire situation, RL agents struggle to infer higher-level semantics (e.g., remaining team count) during early training. Therefore, mid-level summary features—**KO count**, **PP**, and **turn ratio**—act as structured priors that accelerate convergence and reduce reward variance. These variables are partially redundant in an information-theoretic sense but empirically stabilize learning and improve performance. Meanwhile, integer type encoding and the potential feature Φ further compress high-correlation inputs, retaining essence while reducing dimensionality. Despite omitting unobserved or fine-grained details (e.g., unrevealed Pokémon or move metadata), the chosen features remain sufficient for decision-making under an approximate Markov assumption.

Potential feature (Φ). To give the agent a global sense of advantage, a continuous **battle potential** variable $\Phi \in [-1, 1]$ is defined by combining HP difference, type advantage, speed, and weather effects. The idea is inspired by *positional evaluation* in strategic games such as chess and Go [3]: in chess, for example, a player with more active and valuable pieces is said to have the advantage, while a passive or constrained position indicates disadvantage. Similarly, in Pokémon battles, higher HP, faster speed, type superiority, favorable weather, or a successful advantageous switch collectively signal a winning position. The variable Φ therefore quantifies this overall *positional advantage* as a continuous scalar.

$$\Phi = 0.5\,\text{hp\_diff} + 0.25\,\text{type\_adv} + 0.15\,\text{speed\_adv} + 0.10\,\text{weather\_adv}, \quad \Phi \in [-1, 1]. \tag{1}$$

A higher Φ implies stronger overall control, while its change $\Delta\Phi$ measures whether the agent is moving toward victory. In this sense, $\Delta\Phi$ acts as a *dense directional signal* linking intermediate decisions to long-term success. Even without terminal rewards, the agent can sense whether it is improving the position—analogous to gaining material or initiative in chess. This provides smoother feedback, faster convergence, and improved stability during training. Finally, Φ values are clipped to $[-1, 1]$ for numerical consistency.

2.2 Action Space Design

Motivation and design process. At the beginning of training, the action space of Pokémon battles proved prohibitively large and unstable for reinforcement learning. Each turn may involve dozens of context-dependent options—attacking, switching, or triggering special mechanics (e.g., Mega, Z-Move, Dynamax, Terastallize)—that vary dynamically as the battle progresses. Allowing the agent to explore this full space led to divergence and erratic policies. To make learning tractable, the environment was initially designed with a **minimal action set** containing only the four basic attack moves (*move 1–4*). However, early experiments showed that without the ability to switch Pokémon, the agent frequently entered unrecoverable disadvantage states. This prompted an expansion of the action set to include **switch actions** (*slot 1–6*), forming the final 10-discrete-action setup implemented in `_get_action_size()`. This design preserves the core tactical trade-off—**attack** versus **switch**—while removing high-level mechanics that add complexity but little strategic diversity.

Action discretisation. The final implementation fixes the action indices as *0–5: switch to slot 1–6; 6–9: use move 1–4*. This medium-granularity abstraction captures the essential layer of tactical reasoning in Pokémon—offense, defense, and tempo control—while avoiding the combinatorial explosion of fine-grained move mechanics. It provides a clean, interpretable structure that significantly stabilizes training and policy convergence.

Invalid-action fallback. Pokémon battles impose strict legality constraints: a fainted target cannot be switched to, moves with 0 PP or disabled status cannot be used, and attacks are prohibited under forced-switch conditions. To prevent dead trajectories, the environment includes an *automatic fallback mechanism* that detects illegal choices and replaces them with the "best available move," estimated by expected damage $\times$ STAB $\times$ type advantage. This mechanism maintains *stability* (no invalid transitions) and *continuity* (consistent feedback), ensuring that learning remains effective even during random exploration.

Switching heuristics and granularity. Switching is a high-cost decision that can sacrifice tempo and expose the next Pokémon to damage. Random switching in early training introduced large noise and unstable learning. To control this, a simple heuristic was added: switching is allowed only when HP is critically low, the type matchup is severely disadvantageous, or all moves are ineffective; otherwise, the action defaults to attack. This rule preserves strategic realism while suppressing unnecessary randomness. Overall, by limiting decisions to **attack** or **switch**, the agent learns the essential balance between aggression and survival—achieving a stable, interpretable, and computationally efficient control policy.

2.3 Reward Function Design

Motivation and design rationale. In reinforcement learning, rewards define what it means to "do well." In Pokémon battles, however, success emerges not only from the final win/loss outcome but from a sequence of evolving advantages and disadvantages. The core idea of this design was to model how players perceive advantage: every moment in battle can be viewed as either a **winning state** (high potential) or a **losing state** (low potential). Winning states typically feature healthier Pokémon, favorable type matchups, faster speed, or supportive weather, whereas losing states correspond to the opposite conditions—low HP, type disadvantage, and poor tempo control. This qualitative distinction provides an intuitive foundation for quantitative reward shaping.

From state evaluation to potential change. To translate these human-like intuitions into machine feedback, the environment employs a continuous **battle potential** score $\Phi \in [-1, 1]$ (defined earlier in Section 2.1), which estimates the overall positional advantage of the agent. While Φ measures the static strength of a position, its change $\Delta\Phi$ across turns captures the *direction* of progress—whether the agent is moving toward victory or drifting into disadvantage. This differential form is used as a dense intermediate signal that smooths the sparse win/loss reward and stabilizes training. In other words, $\Delta\Phi$ tells the agent not only "what happened" but also "whether it was a step in the right direction."

Hierarchical reward shaping: balancing sparsity and density. Based on this principle, the reward function integrates multiple signals that operate across different temporal and strategic scales. Terminal victory defines the ultimate goal, knockout events represent mid-term progress, and potential change ($\Delta\Phi$) and HP variation deliver dense, continuous feedback. A smaller regularization term on switch quality discourages costly or unnecessary switching. Together, these components form a **multi-scale hierarchy** of feedback signals that range from **sparse** to **dense**, balancing convergence speed with long-term alignment.

$$R_{\text{total}} = 5.0R_{\text{term}} + 2.0R_{\text{KO}} + 2.0R_{\Delta\Phi} + 0.5R_{\text{HP}} + 0.5R_{\text{switch}}. \tag{2}$$

Table 2: Reward components and their design purposes.

Component	Weight	Purpose
Terminal (win/loss)	5.0	Defines the long-term goal; enforces the true success condition ($+1/-1$).
KO event	2.0	Encourages decisive offensive progress; mid-term milestone.
$\Delta\Phi$ (potential change)	2.0	Dense directional signal capturing progress toward victory.
HP delta	0.5	Local frequent feedback for tactical advantage and continuous learning.
Switch quality	0.5	Regularizes high-cost switching behavior to avoid instability.

Balancing sparsity and density. A purely sparse reward (e.g., win/loss) provides clear objectives but insufficient gradient flow, whereas overly dense rewards risk short-sighted behavior or reward hacking. This design strikes a balance through hierarchical composition:

$$\text{Sparse (Win)} \rightarrow \text{Semi-dense (KO)} \rightarrow \text{Dense (HP, } \Delta\Phi\text{)}.$$

Terminal and KO rewards provide long-term strategic direction, while HP and $\Delta\Phi$ offer smoother gradients for exploration and faster convergence. This multi-scale shaping ensures the agent learns both immediate tactics and consistent long-horizon strategies, maintaining interpretability while accelerating convergence.

Alignment and validation. Each component's sign and weight are designed to remain consistent with the final objective: damage to the opponent, successful KOs, and positive $\Delta\Phi$ yield positive rewards, while taking damage or losing Pokémon produces penalties. All terms are clipped to $[-10, 10]$ for numerical stability. To verify the design, each component underwent a *goal-alignment test*: (1) If optimized alone, does it improve the win rate? (2) If overemphasized, does it produce unrealistic behavior? Only components satisfying both criteria were retained or given higher weights. This validation ensures that the overall reward function genuinely drives progress toward winning rather than exploiting unintended loopholes.

2.4 Metrics and Evaluation Criteria

Motivation. To avoid the common pitfall of "rising rewards real learning," evaluation focuses not only on reward magnitude but also on *behavioral stability* and *goal alignment*. An agent is considered to have *truly learned* only when multiple independent indicators show consistent improvement across time.

Evaluation framework. Three complementary categories capture different aspects of learning quality. Table 3 summarizes the core metrics, their types, and purposes.

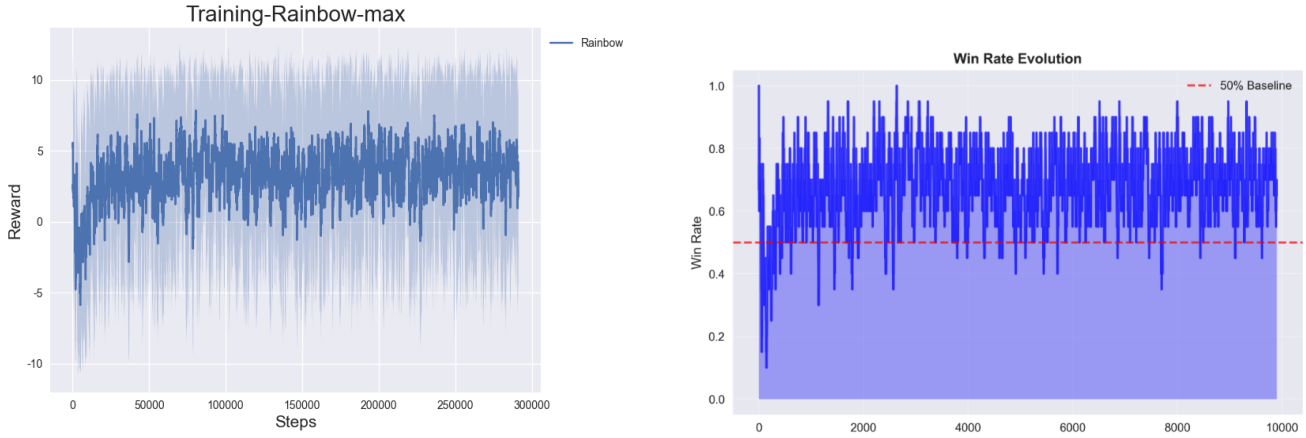
Table 3: Evaluation metrics capturing success, stability, and behavioral alignment.

Metric	Category	Purpose and interpretation
Win rate / Moving winrate	Task success	Ultimate indicator of performance; fraction of battles won.
KO / HP difference	Task success	Measures battle dominance and sustained offensive advantage.
$\Delta\Phi$ (Potential change)	Learning stability	Directional signal of progress; checks whether policy moves toward victory.
Reward / Reward std.	Learning stability	Indicates smoothness of training and gradient flow.
Illegal action rate	Behavioral alignment	Tests rule understanding and environment compliance.
Move/Switch ratio	Behavioral alignment	Evaluates offensive–defensive balance and switching rationality.
Switch reason distribution	Behavioral alignment	Explains why switches occur (low HP, type disadvantage, etc.).

Criteria for success. A policy is regarded as *successfully learned* only when the following hold simultaneously: steady reward rise without oscillation, win rate > 0.6, positive KO/HP difference, mean $\Delta\Phi > 0$ with decreasing variance, and a near-zero illegal-action rate. This indicates both **competence** (ability to win) and **stability** (consistent, rule-aligned strategy improvement).

3 Final Design Results

Overall Training Trend. The Rainbow agent was trained for **23 hours**, covering **291,269 steps** over **9,899 episodes**. As shown in Fig. 1, both training and evaluation rewards show a clear upward trend and gradual stabilization, indicating that the agent learned consistent patterns from the environment. The evaluation reward slightly exceeds the training reward (3.7 vs. 3.4), suggesting mild generalization. The moving winrate stabilizes between **0.68–0.72**, significantly higher than the random baseline (0.5), confirming steady learning progress. However, local oscillations remain due to exploration and reward-weight variability. **Conclusion:** Stable learning achieved, but convergence speed and reward variance remain limiting factors.



(a) Training and Evaluation Reward

(b) Moving Winrate

Figure 1: Reward and winrate curves showing steady convergence.

Battle Dynamics and Strategy Behavior. Fig. 2 presents HP, KO, and tempo-related results. The HP gap between our and opponent teams remains slightly positive (0.2), showing that the agent maintains a mild but consistent advantage. However, the KO difference gradually decreases, and in some runs becomes negative, indicating difficulty in finishing opponents. This imbalance arises because HP-delta rewards dominate over KO-related signals, leading to persistent pressure rather than efficient termination. Speed advantage remains above 80% of turns, showing that the agent learned to seize initiative effectively. **Conclusion:** The model mastered tempo control and continuous pressure, but lacks decisive finishing capability.

Action Distribution and Switching Behavior. Fig. 3 shows that the agent maintains about **70% attack** and **25–30% switching** behavior across episodes. All switch events are triggered under low-HP conditions, demonstrating that heuristic filters successfully prevent random switching. However, post-switch $\Delta\Phi$ improvement remains marginal, suggesting that these defensive switches are reactive rather than strategic. In early versions, the lack of type-advantage awareness limited the model's ability to perform proactive counter-switches. **Conclusion:** The agent displays interpretable and rule-consistent decision patterns, but switching lacks higher-level planning and foresight.

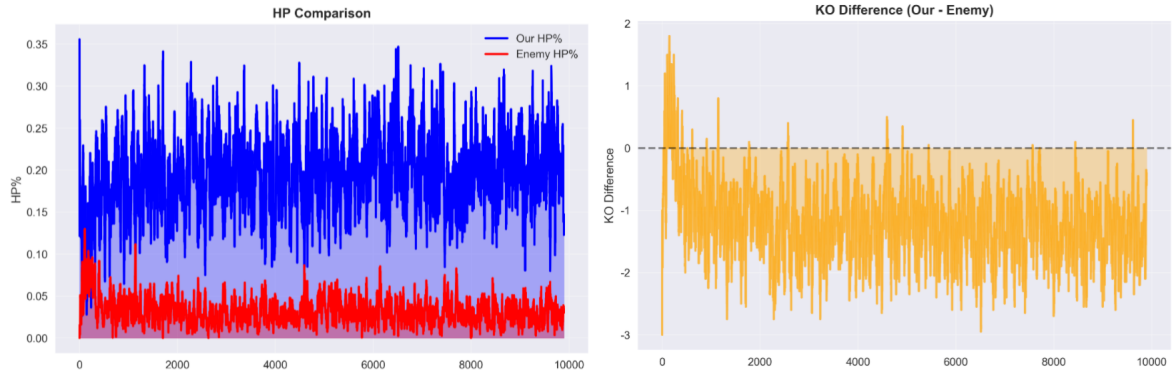


Figure 2: Battle dynamics summarizing HP gap, KO difference, and speed advantage across training.

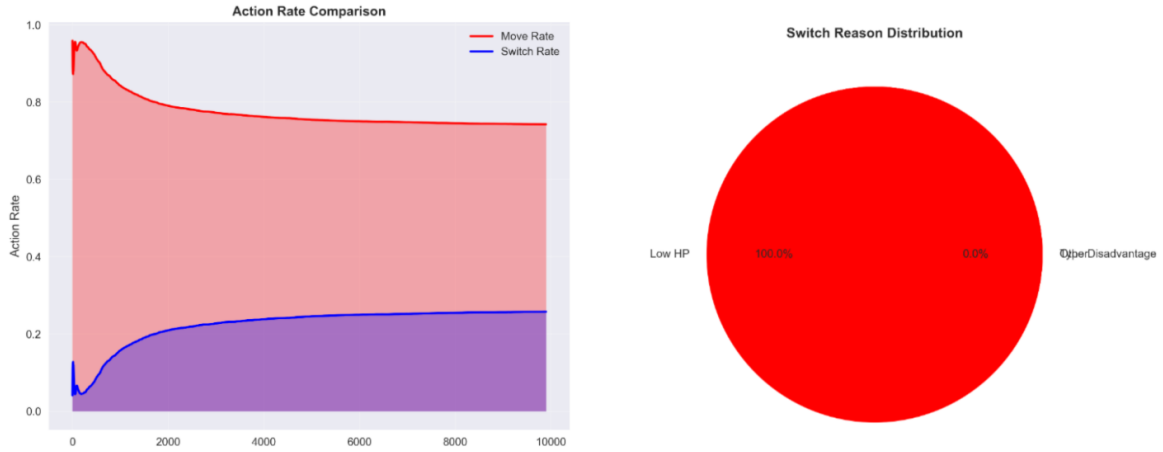


Figure 3: Action balance and switch-motivated behavior.

Potential and Positional Analysis. From Fig. 4, the average potential Φ stays around **0.1–0.2**, indicating a small but persistent advantage throughout training. Mean $\Delta\Phi$ is approximately **+0.011**, showing gradual improvement per turn, yet high variance reveals unstable decision quality in complex states. The $\Delta\Phi$ distribution is right-skewed with peaks near zero, implying limited momentum toward sustained advantage. Overall, the potential-based shaping successfully teaches directional improvement but struggles to maintain consistency over long horizons. **Conclusion:** The shaping term provides dense feedback but reaches diminishing returns as the battle length increases.

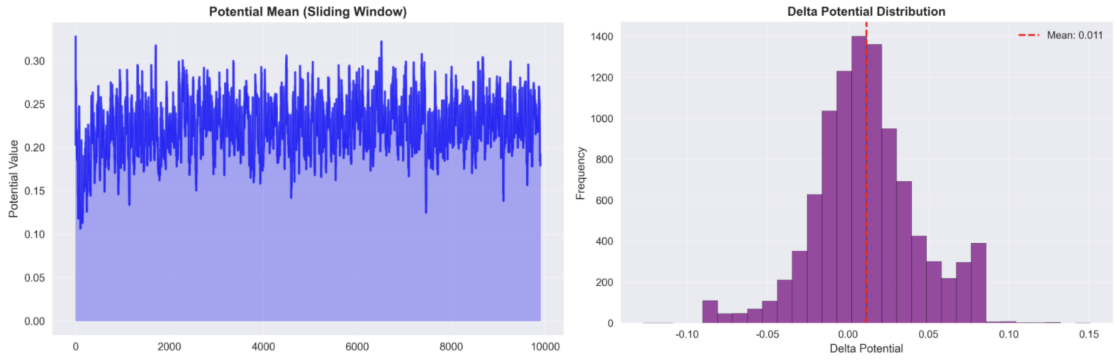


Figure 4: Potential stability and directional improvement analysis.

Overall Performance Summary and Discussion.

Overall, the Rainbow model demonstrates strong *reward alignment*, robust *stability*, and clear *rule awareness*. It recognizes advantageous moments and maintains tempo effectively, yet lacks decisive endgame execution. The potential-based shaping accelerates learning and yields interpretable intermediate signals, but also weakens long-horizon planning. **In summary:** The agent can *"recognize direction"*—knowing when it is winning or losing— but cannot yet *"understand the objective"*—how to convert advantage into victory. This insight highlights both the power and the limitation of potential-driven shaping in multi-step competitive RL.

Table 4: Final evaluation results and key indicators.

Metric	Value	Interpretation
Final Winrate (Eval)	**0.717**	Clearly above baseline; stable generalization.
Average Reward	3.4 / 3.7 (train/eval)	Reward shaping directionally consistent with victory.
KO Difference	$\approx$0.3	Lack of finishing efficiency; preference for pressure play.
Mean Potential (Φ)	0.18	Slight but unsustained positional advantage.
Mean $\Delta\Phi$	+0.011	Small but positive per-turn progress.
Illegal Action Rate	$<$0.5%	Strong rule compliance and environment stability.

4 Discussion and Reflection

Defining the world matters more than tuning parameters. Early experiments showed that performance was driven less by hyperparameter search than by how the environment was *defined*. A clear and compact world definition—observable states, valid actions, and goal-aligned rewards—determines what the agent can meaningfully learn. When the world is well-posed, even simple algorithms converge; when it is noisy or redundant, extra complexity only amplifies instability.

Reward–success alignment. A good reward function must be both learnable and directional. In this project, dense rewards (HP, potential shaping) accelerated learning but also shifted the agent's focus toward maintaining high HP rather than finishing battles. Sparse signals (KO, win) were slower but preserved strategic intent. This trade-off highlights that reward shaping should *guide*, not *replace*, the true objective. The most effective signals were those where local improvement implied global progress—e.g., $\Delta\Phi > 0$ coinciding with higher winrate.

KO vs. HP: sparse–dense complementarity. HP feedback appears every step, stabilizing gradients but promoting cautious play; KO events, though rare, convey the decisive long-term goal. An ideal design combines them hierarchically: dense rewards for exploration and consistency, sparse rewards for strategy and closure. Future work could use adaptive weighting, where KO/terminal rewards gradually dominate as training stabilizes.

Potential shaping as a bridge. The potential feature Φ successfully connected low-level state feedback to the global notion of advantage. However, as a single scalar it cannot express the structure of multi-factor interactions—type, speed, and weather influence advantage differently. Extending Φ into a multi-dimensional potential vector could yield richer, yet still compact, guidance for decision making.

Stability as the true indicator of understanding. Beyond numerical rewards, genuine learning is reflected in consistent and interpretable behavior. When similar states yield similar actions and the variance of $\Delta\Phi$ decreases, the policy can be said to "understand" the game's dynamics rather than merely memorizing outcomes. Stability and alignment, not raw reward growth, are therefore the clearest signs of competence.

Future directions. Further improvements could involve adaptive reward weighting, phase-based curriculum training (opening $\rightarrow$ control $\rightarrow$ closure), and multi-agent self-play for richer diversity. The key challenge ahead is to couple fast, dense learning signals with long-horizon strategic reasoning—achieving an agent that not only *knows when it is winning*, but also *knows how to win*.

References

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016. URL `https://arxiv.org/abs/1606.06565`.

[2] Andrew Y. Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML)*, pages 278–287, 1999.

[3] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018. doi: 10.1126/science.aar6404.

[4] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018. URL `http://incompleteideas.net/book/the-book-2nd.html`.