Expert-Guided Training of a Neural Association Rules Croquet Simulator

Michael Kernaghan
michaelkernaghan@ecadlabs.com

January 10, 2026

Abstract

We present a deep reinforcement learning approach to Association Croquet, a long-horizon physical game with continuous state, sparse rewards, and complex expert strategy. Unlike recent successes in pure self-play for board games, croquet requires structured tactical knowledge to learn effective break construction. We combine a dueling Deep Q-Network with expert-derived reward shaping and action-space augmentation informed by elite-level croquet theory.

A key finding is that naive reward designs—such as incentivizing ball clustering—produce fundamentally incorrect play, while proper break construction requires explicit support for pioneer placement during croquet strokes. We show that this capability was absent from the original action space and integrate expert heuristics to address this limitation.

Our results demonstrate substantial improvements in break-building behavior and highlight broader lessons about reward engineering, action-space completeness, and the practical role of expert knowledge in reinforcement learning for complex physical games under limited computational resources.

1 Introduction

Association Croquet is a two-player lawn game with a rich strategic tradition dating to the 1860s. Unlike its casual garden-party image, competitive croquet at the elite level—exemplified by the MacRobertson Shield, the sport's premier international competition—involves deep tactical reasoning comparable to chess [Wylie, 1985].

The game presents several challenges for artificial intelligence:

- **Long-horizon planning**: A player must navigate 26 hoops (including the peg-out) with four balls, requiring planning over potentially hundreds of shots.
- **Sparse terminal rewards**: Games can last hours with only win/loss as natural feedback.
- **Complex state space**: Ball positions are continuous, hoop progress is discrete, and the interaction between offensive and defensive considerations creates a large effective state space.
- **Break construction**: The signature skill of croquet—running a "break" where one scores multiple hoops in a single turn—requires precise ball positioning that differs fundamentally from intuitive notions of "keeping balls together."

This paper describes our approach to training a croquet-playing agent using deep reinforcement learning augmented with expert knowledge. Our key contributions are:

1. A croquet simulation environment suitable for reinforcement learning

2. An automated pipeline for extracting tactical patterns from video transcripts and structured game notation
3. Expert-derived reward shaping based on five years of elite tournament data (2020–2025)
4. Analysis of why naive reward functions (e.g., rewarding ball clustering) produce suboptimal play
5. A dueling DQN architecture with n-step returns adapted for croquet's long horizons
6. Discussion of the expert-guided vs. pure self-play trade-off in sports AI

Positioning. This paper is not intended as a benchmark-setting result or a claim of optimal play in Association Croquet. Rather, it serves as a focused case study demonstrating how reward shaping, expert priors, and action-space design interact in long-horizon physical games. Our goal is to surface failure modes, design insights, and practical trade-offs relevant to reinforcement learning systems operating under realistic data and compute constraints.

2 Background

2.1 Association Croquet Rules

Association Croquet is played with four balls—blue and black versus red and yellow—on a standard lawn with six hoops arranged in a specific pattern. Each ball must run all six hoops twice (in prescribed directions) plus hit the center peg, for 13 points per ball and 26 points total to win.

The key mechanic is the *croquet stroke*: when a player's ball hits another ball (a "roquet"), they earn two bonus strokes—first a croquet stroke where both balls move, then a continuation stroke. Skilled players chain these interactions into "breaks" scoring many hoops consecutively.

2.2 The Four-Ball Break

The canonical technique in competitive croquet is the *four-ball break*, where a player uses all four balls in coordinated positions:

- **Striker's ball**: The ball being played
- **Pilot ball**: Positioned near the current hoop to assist running it
- **Pioneer**: Placed ahead at the next hoop
- **Pivot**: Centrally located to provide access to other balls

Critically, proper break construction requires balls to be *spread out* with pioneers placed *ahead* at upcoming hoops. This contradicts the naive intuition that keeping balls clustered together is advantageous. Clustering is only appropriate in specific situations: setting a defensive "leave" at the end of a break, or positioning for the final peg-out sequence.

2.3 Deep Q-Networks

Deep Q-Networks [Mnih et al., 2015] approximate the action-value function $Q(s, a)$ with a neural network. The network is trained to minimize the temporal difference error:

$$\mathcal{L} = \mathbb{E}\left[\left(r + \gamma \max_{a'} Q(s', a'; \theta^-) - Q(s, a; \theta)\right)^2\right] \tag{1}$$

where θ^- represents a periodically-updated target network.

2.3.1 Dueling Architecture

The dueling DQN architecture [Wang et al., 2016] separates the value function $V(s)$ from the advantage function $A(s, a)$:

$$Q(s, a) = V(s) + A(s, a) - \frac{1}{|\mathcal{A}|} \sum_{a'} A(s, a') \tag{2}$$

This separation helps the network learn which states are valuable independent of action choice, particularly useful in croquet where many states have clear strategic value regardless of the specific shot selected.

2.3.2 N-Step Returns

Rather than single-step TD updates, n-step returns provide better credit assignment for delayed rewards:

$$G_t^{(n)} = \sum_{k=0}^{n-1} \gamma^k r_{t+k+1} + \gamma^n \max_a Q(s_{t+n}, a) \tag{3}$$

We use $n = 3$, balancing the bias-variance trade-off for croquet's medium-horizon tactical sequences.

3 Related Work

3.1 Game-Playing AI

Deep reinforcement learning has achieved superhuman performance in numerous games. AlphaGo [Silver et al., 2016] combined Monte Carlo Tree Search with deep networks, initially using expert game records before AlphaZero [Silver et al., 2018] demonstrated pure self-play learning. These approaches require substantial computational resources—thousands of TPUs training for days.

For board games with perfect information, pure self-play is theoretically sufficient. However, physical sports simulations introduce additional complexity: continuous state spaces, physics modeling, and the need to learn basic motor skills alongside strategy.

3.2 Sports Analytics and AI

AI applications in sports have primarily focused on analytics—player tracking, outcome prediction, and tactical analysis [Decroos et al., 2019]. Fewer works address learning to play physical sports from scratch, likely due to the combined challenges of physical simulation and strategic depth.

3.3 Reward Shaping

Reward shaping [Ng et al., 1999] accelerates learning by providing intermediate feedback. Potential-based shaping preserves optimal policies while guiding exploration. Our approach uses non-potential-based shaping derived from expert knowledge, accepting the trade-off of potentially suboptimal but practically effective policies.

4 Methods

4.1 Simulation Environment

Our croquet simulator implements the full Association Croquet ruleset including:

- Realistic ball physics with collision detection
- All stroke types (roquet, croquet, continuation)
- Hoop running mechanics
- Turn structure with proper deadness rules
- Fault detection and wiring calculations

The state representation includes:

- Ball positions (8 continuous values)
- Hoop progress for each ball (4 integers, 0-13)
- Current striker and ball in hand
- Deadness state (which balls can be roqueted)
- Turn phase (approach, croquet, continuation)

Actions are discretized into shot types (roquet attempts, hoop runs, position plays) with parameterized strength and direction.

4.2 Expert Prior Extraction

We analyzed commentary from MacRobertson Shield matches and the 2025 WCF World Championship, extracting tactical patterns from elite players including Robert Fulford, Reg Bamford, Mark Avery, and other world champions. Key patterns encoded:

- **Pioneer placement priority**: Strong reward for placing balls ahead at upcoming hoops
- **Break continuation**: Bonus for running hoops while maintaining break structure
- **Pivot utilization**: Reward for maintaining central pivot ball access
- **Leave construction**: Appropriate clustering only at break termination

4.2.1 Data Collection Pipeline

We developed an automated pipeline to collect and process expert game data from multiple sources:

Video Commentary Transcripts. Using yt-dlp, we extracted auto-generated transcripts from 18 official 2025 WCF World Championship live stream videos, totaling 687,417 words. These transcripts capture real-time expert commentary describing tactical decisions, shot selection, and positional play.

Structured Game Notation. From CroquetScores.com, we obtained detailed turn-by-turn game transcriptions using standard croquet notation. These include precise positional information (e.g., "2 yards NW of hoop 5", "Blue 10y N of IV"), shot types, and outcomes. Data sources span four years of elite competition: the 2025 WCF World Championship semifinal (Fletcher v Avery), the complete 2024 Croquet England Open Championships (Round of 16 through Final), the 2023 WCF World Championship (Quarterfinals through Final, including the Essick v Fulford final), the 2023 AC Open Singles Championship (Quarterfinals through Final, featuring the Avery v Death final), and the 2022 CA AC Open Singles Championship (Quarterfinals

through Final, featuring the Death v Bamford final with sextuples and self-peg-out tactics). All matches were annotated by Andrew Gregory. This yielded 395 fully-annotated turns across expert games, capturing advanced techniques including sextuple peels, quintuple peels, quadruple peels, TPO (triple peel on opponent) plays, delayed triple strategies, and supershot opening positions.

Pattern Extraction. We developed custom parsers to identify tactical patterns in both data sources. Table 1 summarizes the extracted patterns.

Table 1: Tactical patterns extracted from expert transcripts

Pattern Type	Count	Avg. Reward
Peel sequence	95	1.76
Break assembly	82	1.58
Rush control	76	1.82
Break continuation	56	2.26
Rush to hoop	25	0.60
Leave setting	18	0.17
Pioneer placement	4	0.60
Take-off	2	0.50

The relatively low count for pioneer placement patterns (4 instances) despite its tactical importance suggests that commentators often assume this knowledge, discussing pioneers implicitly rather than explicitly naming the technique. This highlights a limitation of transcript-based learning: expert knowledge is often tacit.

4.2.2 Training Data Statistics

Our final dataset comprises:

- 687,417 words from video commentary (2.1% tactical density)
- 437 structured game turns with precise notation
- 510 unique training examples across 15 pattern categories
- Expert game demonstrations from five tournament years (2020–2025)

Expert sources include:

- Keith Aiton's "Expert Croquet Tactics" [Aiton, 2009]
- 2025 WCF World Championship video commentary (18 sessions)
- 2025 WCF World Championship semifinal notation (Andrew Gregory)
- 2024 Croquet England Open Championships—Round of 16 through Final (Andrew Gregory)
- 2023 WCF World Championship—Quarterfinals through Final (Andrew Gregory)
- 2023 AC Open Singles Championship—Quarterfinals through Final (Andrew Gregory)
- 2022 CA AC Open Singles Championship—Quarterfinals through Final (Andrew Gregory)
- 2020 WCF World Championship—Knockout through Final (Melbourne, Australia)
- Oxford Croquet archival analysis

4.3 Reward Function Design

Our reward function combines sparse game rewards with shaped tactical rewards:

$$r_t = r_{\text{game}} + r_{\text{tactical}} \tag{4}$$

4.3.1 Game Rewards

- Hoop run: +10
- Peg-out: +20
- Game win: +100
- Game loss: −100

4.3.2 Tactical Rewards (Expert-Derived)

- Pioneer created at next hoop: +4.0
- Pioneer created at next-but-one hoop: +2.5
- Break execution with pioneers in place: +5.0
- Successful roquet: +0.8
- Rush executed: +1.0

4.3.3 Removed "Hack" Penalties

Initial versions included several artificial penalties that we subsequently removed based on the principle that proper tactical rewards should make good behavior emerge naturally:

- **Time penalty**: Removed—efficient play should emerge from tactical rewards
- **Rover penalty**: Removed—peg-out incentives sufficient
- **Defensive action penalty**: Removed—defensive play is sometimes correct
- **Cluster reward**: Removed—was teaching improper bunching behavior

4.4 Why Clustering is Wrong

An early version rewarded "cluster quality"—keeping balls close together. This produced agents that bunched all four balls near the first hoop, unable to progress.

The error reflects a fundamental misunderstanding of croquet tactics. In a proper four-ball break:

1. The pilot assists at the current hoop
2. The pioneer waits at the *next* hoop
3. Another pioneer may be at the hoop after that
4. The pivot provides central access

Balls must be spread across the lawn with specific positional relationships. Clustering only occurs when:

- Setting a defensive leave (end of break)
- Preparing for rover/peg-out (endgame)

This illustrates why domain expertise matters: naive reward engineering can encode precisely the wrong behavior.

4.5 Network Architecture

Our dueling DQN uses:

- Input layer: State features (positions, progress, phase)
- Hidden layers: 256-128-64 units with ReLU activation
- Dueling streams: Separate value and advantage heads
- Output: Q-values for discretized action space

Training hyperparameters:

- Learning rate: 10^{-4}
- Discount factor γ: 0.99
- Replay buffer size: 100,000 transitions
- Batch size: 64
- Target network update: Every 1,000 steps
- N-step returns: 3
- Exploration: ϵ-greedy with decay

5 Experiments

5.1 Training Protocol

We train agents for 2,000 episodes with evaluation every 100 episodes. Checkpoints are saved every 500 episodes for analysis.

Training configurations tested:

- Baseline DQN (no expert shaping)
- DQN with expert priors
- Dueling DQN with expert priors
- Dueling DQN with expert priors and 3-step returns

5.2 Evaluation Metrics

- **Win rate**: Against baseline opponent
- **Average hoops per turn**: Measure of break-building
- **Pioneer placement rate**: Frequency of proper positioning
- **Break length distribution**: Histogram of consecutive hoops

6 Results

6.1 Training Performance

Training with expert-derived rewards and Aiton's croquet stroke mechanics produced the following key performance indicators:

Table 2: Training results with Aiton croquet stroke integration

Metric	Value
Pioneer placement rate	33%
Pilot ball positioning	30%
Rush execution rate	7%
Hoops per game	15.7
Break balls utilized	2.5
Cluster coefficient	0.45
Average reward	19,239

6.1.1 Metric Definitions

To clarify the tactical KPIs reported above:

- **Pioneer placement rate**: Percentage of croquet strokes where a ball was placed within 5 yards of an upcoming hoop (not the current hoop). Expert human play typically achieves 60–80% in controlled breaks.
- **Pilot ball positioning**: Percentage of turns where a ball was positioned near the current target hoop to assist the approach. Expert range: 70–90%.
- **Rush execution rate**: Percentage of roquets that resulted in the roqueted ball moving toward a tactically useful position. Expert range: 40–60%.
- **Break balls utilized**: Average number of the four balls actively used in break construction per turn (range 1–4). A proper four-ball break uses all 4; our agent averages 2.5.
- **Cluster coefficient**: Average pairwise distance between balls normalized to court diagonal, where 0 = all balls coincident and 1 = maximally spread. Values of 0.4–0.6 indicate appropriate tactical spread rather than bunching.
- **Average reward**: Cumulative shaped reward per episode, combining game outcomes (+100 win, -100 loss) with tactical bonuses (hoop runs, pioneer placement, roquets).

Our agent's metrics (33% pioneer, 30% pilot, 2.5 break balls) indicate competent but not expert-level break construction. The agent successfully avoids the naive bunching behavior that plagued early reward designs.

6.2 Key Finding: Action Space Limitation

A critical discovery during development was that the neural network's action space—comprising 8 high-level intents (hoop run, roquet attempts, approach, defensive, peg-out)—contained no action for placing pioneers during croquet strokes. After a successful roquet, the agent could not control where to send the croqueted ball.

The solution integrated Aiton's tactical placement heuristics directly into the training loop: when the rules engine detected a croquet-required state, we invoked the AIController's croquet shot selection method, which implements Aiton's teachings on pioneer placement (positioning balls 3–4 yards in front of the next hoop).

6.2.1 Characterizing the Hybrid Approach

This hybrid architecture warrants explicit framing: **the neural network learns *when* and *why* to place pioneers, but not *how* to execute the placement physics**. The agent decides to roquet a particular ball (learned behavior), and the expert heuristic then determines where that ball should be sent during the subsequent croquet stroke (programmed behavior).

This division of labor is defensible for several reasons:

- The "how" of croquet stroke execution is well-understood physics with optimal solutions documented in coaching literature
- Learning low-level motor control would require orders of magnitude more training without strategic benefit
- The agent still learns the strategically meaningful decisions: which ball to roquet, when to continue a break versus set a leave, and how to recover from poor positions

This approach parallels how human players learn: beginners are taught stroke mechanics explicitly ("place the pioneer 3–4 yards in front of the next hoop"), then develop strategic judgment through experience.

6.3 Before and After Comparison

Without Aiton croquet stroke integration, the agent achieved:

- Pioneer placement: 0% (unable to learn)
- Hoops per game: 3.4

With integration:

- Pioneer placement: 33%
- Hoops per game: 15.7

This 4.6× improvement in hoops per game demonstrates that proper break construction—spreading balls across the lawn with pioneers at upcoming hoops—is essential for competitive croquet play.

6.4 Baseline Comparisons

To contextualize our results, we compare against several baseline agents:

Table 3: Comparison with baseline agents

Agent	Hoops/Game	Pioneer %	Win vs Random
Random-legal	1.2	8%	50%
Clustering-reward DQN	3.1	0%	52%
Base DQN (no expert shaping)	3.4	0%	55%
Expert-shaped DQN (ours)	15.7	33%	78%

- **Random-legal agent**: Selects uniformly from legal actions. Achieves 1.2 hoops/game through chance.
- **Clustering-reward DQN**: Trained with naive clustering bonus (Section 4.4). Learns to bunch balls near hoop 1, unable to progress. Demonstrates that incorrect reward shaping produces worse results than no shaping.
- **Base DQN**: Identical architecture, no expert rewards, no Aiton croquet integration. Action space limitation prevents learning pioneer placement.
- **Expert-shaped DQN (ours)**: Full system with expert rewards and Aiton croquet strokes.

All DQN variants share identical architecture (dueling, 3-step returns) and hyperparameters. The clustering-reward agent serves as a "negative baseline" demonstrating how domain-naive reward engineering can encode precisely wrong behavior.

6.5 Training Convergence

Current results reflect early-stage training. Performance metrics were still improving at training termination, suggesting additional episodes would yield further gains. Extended training runs are ongoing; we report preliminary results to establish the validity of the approach rather than claim convergence to optimal play.

7 Discussion

7.1 Expert Guidance vs. Pure Self-Play

Our approach explicitly incorporates expert knowledge through reward shaping. This contrasts with AlphaZero's pure self-play paradigm, raising the question: which approach is better?

Arguments for expert guidance:

- Dramatically reduced compute requirements
- Encodes centuries of accumulated tactical wisdom
- Avoids rediscovering known strategies from scratch
- Practical for resource-constrained research

Arguments for pure self-play:

- May discover novel strategies beyond expert knowledge
- No dependency on domain expertise or transcriptions
- Theoretically cleaner—learns from game rules alone
- Demonstrated superhuman performance in other games

For croquet specifically, we argue that expert-guided learning is currently more practical:

1. Croquet's physical simulation adds complexity beyond abstract board games
2. The state space (continuous positions, 26-hoop progression, complex interactions) would require massive self-play exploration
3. Expert knowledge in croquet is well-documented and encodes genuine tactical insights

4. Our computational resources are limited compared to DeepMind's infrastructure

A potential hybrid approach: use expert shaping for initial training, then gradually reduce expert influence while increasing reliance on game outcomes. This could combine the bootstrap efficiency of expert guidance with the potential for self-play to discover improvements.

7.2 Transcription Dependency and Data Authority

A valid concern with our approach is dependency on expert game transcriptions, which are relatively scarce for croquet. We distinguish between two data sources with different levels of authority:

Video commentary transcripts capture what experts *verbalize*—often high-level strategic observations, emotional reactions, and assumptions about viewer knowledge. The low count for explicit pioneer placement mentions (4 instances in Table 1) reflects that commentators assume this foundational technique, discussing it implicitly rather than naming it. Transcripts are valuable for identifying *what experts find noteworthy*, not necessarily what they consistently do.

Structured game notation (e.g., CroquetScores annotations) provides authoritative positional data: "Blue 3 yards NW of hoop 4," "Red approaches from boundary." This notation captures *what actually happened* regardless of commentary. Our 437 annotated turns from structured notation are the more reliable tactical source.

Mitigating factors for limited data:

- Core tactical patterns are well-documented in coaching literature (especially Aiton and Wylie)
- Transcriptions identify patterns; the reward function encodes general principles
- Even limited expert data provides substantial signal compared to pure random exploration

Future work could explore:

- Semi-supervised approaches using unlabeled game records
- Transfer learning from related domains
- Curriculum learning to reduce expert dependency over time

7.3 Limitations

- Simulation fidelity: Real croquet involves factors (lawn conditions, equipment variation) not modeled
- Action discretization: Continuous shot parameters are binned, potentially limiting expressiveness
- Evaluation difficulty: No established AI baselines exist for croquet comparison

7.4 Lessons for Reinforcement Learning in Physical Games

This case study surfaces several insights applicable beyond croquet:

- **Reward shaping can encode incorrect domain theory.** Naive rewards based on intuitive notions (ball clustering) produced agents that learned precisely the wrong behavior. Domain expertise is not optional decoration—it determines whether shaped rewards help or harm.

- **Action-space completeness is a prerequisite for learning.** Our agent could not learn pioneer placement because no action existed to control croquet stroke destinations. Incomplete action spaces create hard ceilings on learnable behavior that no amount of training can overcome.
- **Expert heuristics can unblock learning rather than replace it.** Integrating Aiton's placement rules did not reduce the agent to a scripted system—it enabled learning of higher-level strategic decisions that were previously inaccessible.
- **Transcript-based supervision captures explicit but not tacit knowledge.** Commentary reveals what experts find noteworthy, not what they consistently do. Structured notation provides more authoritative behavioral data.
- **Hybrid systems are pragmatic under compute constraints.** Pure self-play may be theoretically elegant, but expert-guided approaches offer practical paths for complex domains where computational resources are limited.

8 Conclusion

We presented a deep reinforcement learning approach to Association Croquet that combines DQN with expert-derived reward shaping. Our key finding—that naive clustering rewards produce fundamentally incorrect play while pioneer-focused rewards enable proper break construction—illustrates the importance of domain expertise in reward engineering.

The tension between expert-guided and pure self-play learning reflects broader questions in AI research. For complex physical sports with limited computational budgets, expert guidance provides a practical path forward. Whether such approaches can eventually match or exceed pure self-play systems remains an open question.

8.1 Future Work

- Extended training with larger episode counts
- Monte Carlo Tree Search integration (MCTS + DQN)
- Multi-agent training with self-play
- Gradual expert-influence reduction experiments
- Real-world deployment on robotic croquet system

Acknowledgments

We acknowledge the croquet community for maintaining historical records and tactical analyses, particularly the Oxford Croquet archive and MacRobertson Shield documentation.

References

Keith Aiton. Expert Croquet Tactics. Croquet Association, 2009.

Tom Decroos, Lotte Bransen, Jan Van Haaren, and Jesse Davis. Actions speak louder than goals: Valuing player actions in soccer. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, volume 99, pages 278–287, 1999.

David Silver, Aja Huang, Chris J Maddison, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

David Silver, Thomas Hubert, Julian Schrittwieser, et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144, 2018.

Ziyu Wang, Tom Schaul, Matteo Hessel, et al. Dueling network architectures for deep reinforcement learning. In *International Conference on Machine Learning*, pages 1995–2003, 2016.

Keith Wylie. Expert Croquet Tactics. EP Publishing, 1985.