Reinforcement Learning Agent for the Mill Game

A Comparative Study of Training Against Multiple Baselines

Marija Ćetković (89252106) & Dmytro Tupkalenko (89252101)

University of Primorska, FAMNIT

January 2026

Introduction & Problem Statement

The Problem

Nine Men's Morris presents two key challenges for Reinforcement Learning:

1. **Sparse legal moves** — Action space of $25^3 = 15,625$ with mostly illegal actions
2. **Three distinct game phases** — Placement $\rightarrow$ Movement $\rightarrow$ Flying

Research Questions

Q1: Can RL agents match classical Minimax performance?

Q2: How does training regime affect learning efficiency and final performance?

Methods

PPO with Action Masking

- **Architecture:** Actor-critic with 2 hidden layers (256 neurons each), tanh activation
- **Action space:** Integer-encoded multi-discrete space ($25^3 = 15,625$ actions)
- **Action Masking:** Mask illegal actions $\rightarrow$ focus on strategy, not rule learning
- **Hyperparameters:** 512 steps/update, $\gamma = 0.99$, $\lambda = 0.95$, 3 epochs, batch=64

Three Training Regimes

Random Agent

- Baseline
- Uniform selection from legal moves

Weak Minimax

- Depth = 1
- 50% random moves

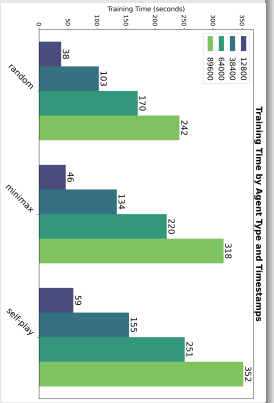
Self-Play

- Frozen policy
- Updated every 10% of training

Results

Evaluation: 1,000 games vs Random, Apprentice (d=1), Adventurer (d=2), Knight (d=3), 50% random moves

Trained Against	Steps	Random	Apprentice	Adventurer	Knight
Random	12800	74.7 : 52.8	41.9 : 43.6	26.1 : 45.2	13.5 : 41.6
	38400	99.2 : 30.5	86.8 : 34.7	57.1 : 43.6	52.4 : 40.6
	64000	99.3 : 28.6	89.8 : 34.3	62.3 : 42.2	54.6 : 40.8
	89600	99.9 : 27.7	91.0 : 33.8	65.3 : 42.5	57.6 : 41.0
Minimax	12800	82.1 : 45.4	75.9 : 37.4	52.7 : 44.9	40.0 : 43.5
	38400	99.3 : 28.6	96.8 : 26.8	80.4 : 37.9	68.7 : 38.2
	64000	99.4 : 29.4	97.7 : 27.1	75.5 : 38.3	66.4 : 38.9
	89600	99.9 : 30.4	96.7 : 27.4	81.0 : 38.4	69.3 : 37.0
Selfplay	12800	81.1 : 48.0	33.4 : 44.0	16.4 : 44.8	11.3 : 42.0
	38400	98.7 : 30.4	73.8 : 33.8	60.8 : 41.8	51.0 : 41.3
	64000	98.8 : 29.0	84.4 : 32.7	71.8 : 38.5	57.6 : 38.0
	89600	99.8 : 29.1	79.9 : 32.3	75.4 : 37.0	63.8 : 38.4



Conclusions

Key Findings

- **RL matches classical methods:** PPO with action masking achieves competitive performance vs Minimax
- **Opponent choice matters:** Structured Minimax opponents yield superior learning efficiency
- **Self-play shows promise:** Strongest late-stage trajectory suggests eventual superiority with extended training

Code: github.com/tupkalenkodi/mill_rl_agent