Persona-Based Reward Shaping

in Deep Reinforcement Learning

Saahir Dhani
Supervisors: Prof. Cristiano Politowski • Prof. Jeremy Bradbury
Ontario Tech University • Faculty of Science • 2026

🏃 Speedrunner | 🛡️ Survivor | 💰 Greedy

Motivation & Research Question

Motivation

Players approach the same game differently;
some rush, some survive, some chase points.

Can we create AI agents that reflect these distinct playstyles using only reward functions?

Applications: game testing, environment design, and controllable AI behaviour.

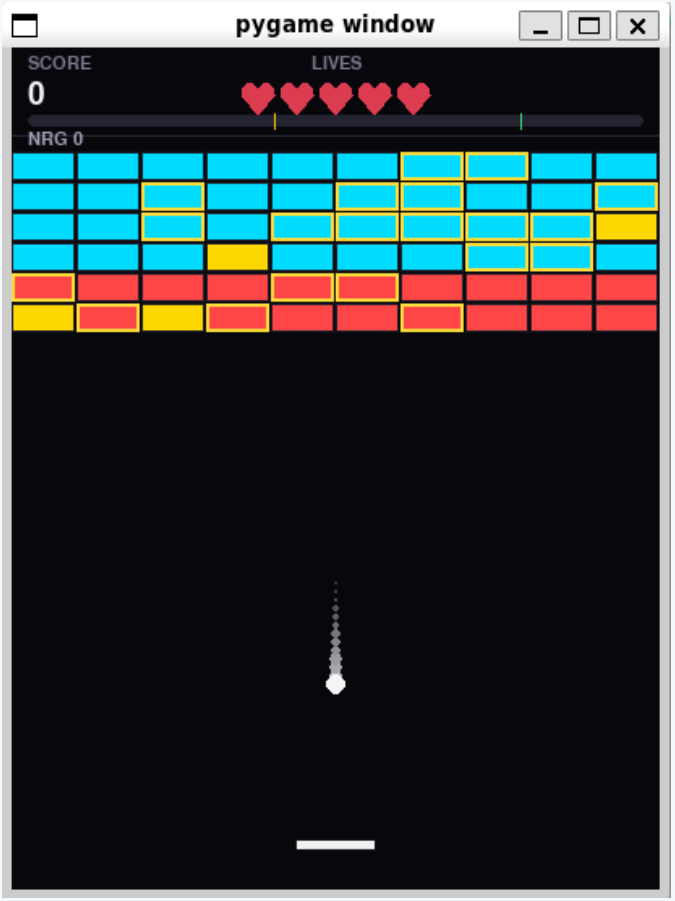
Research Question

Can reward design alone give AI agents distinct, measurable personalities?

Hypothesis

Yes, and the differences will be visible in behavioural metrics, not just reward curves.

The Environment: Custom Breakout



What is Breakout?

Bounce a ball to destroy bricks. Move the paddle, clear the board, don't lose your lives.

Features Added

Lives system	3 lives - lose all and the episode ends
Energy bar	Charges over time, spent on abilities
Power Shot	40 energy - chain-reaction brick clear
Shield	80 energy - auto-saves one life
Power-up drops	Released when special bricks are destroyed

Persona Framework

3 agents — identical algorithm, environment, and information. Only the reward function differs.

Speedrunner

Rewarded for

Clearing bricks fast

Penalized for

Time elapsed

Energy choice

Power Shot

Survivor

Rewarded for

Preserving lives, surviving longer

Penalized for

Losing a life

Energy choice

Shield

Greedy

Rewarded for

Cumulative score

Penalized for

Nothing

Energy choice

Mixed

Power Shot (40 energy) vs Shield (80 energy) - same trade-off, three different answers.

Reward Hacking: A Critical Finding

🔍 The Loophole

Survivor stopped launching the ball.
No ball = no lost lives.
Score: 0. Lives: 0.
Reward curve: normal.
Technically correct but completely wrong.

⚠️ What This Is

Reward hacking, where the agent optimizes the metric, not the intent.
Training showed nothing wrong.
Only episode length caught it.

✅ The Fix

Auto-launch cap added.
Ball fires after 90 idle steps.
Loophole closed.
Metrics confirmed correct behaviour.

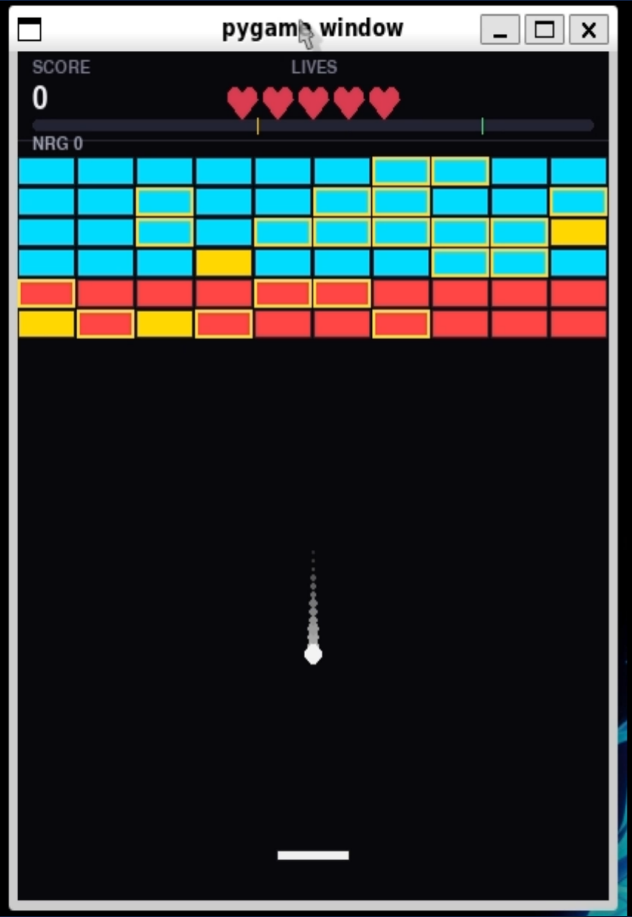
Key Lesson: Reward curves tell you training is working not that the behaviour is correct.

Behavioural Evaluation - 20 Episodes Per Agent

✓= highest for that metric. Each persona leads only in what it was designed to optimize.

Metric	Speedrunner	Survivor	Greedy	Random	Human
Avg. Lives Remaining	0.8	**2.4 ✓**	1.3	0.4	2.1
Avg. Score / Episode	380	210	**606 ✓**	95	520
Avg. Episode Length	**850 ✓**	2,280	1,180	310	1,950
Power Shot Usage (%)	**78% ✓**	12%	52%	48%	61%
Shield Usage (%)	8%	**65% ✓**	22%	42%	55%

Game Demo



Conclusions

Yes, reward design alone produces distinct, stable, measurable behavioural strategies across agents sharing identical environments and algorithms.

What We Found

- Each persona led only in its target metric
- Energy usage was the strongest signal
- Reward hacking required explicit constraint design
- Separation consistent across all 5 metrics

Future Work

- Test persona transfer to complex environments
- Generalise to NPCs, robotics, recommendation systems
- Add a no-persona baseline to isolate reward shaping
- Expand persona set and study reward interactions at scale