DigiSoup: A Zero-Training Entropy-Driven Agent Beats Trained Reinforcement Learning on Multi-Agent Social Dilemmas

Matthew Fearne
Independent Researcher
mrfearne@gmail.com
ORCID: 0009-0005-3986-4927
https://github.com/matthewfearne/digisoup

February 2026

Abstract

We present DigiSoup, a zero-training agent that uses thermodynamic perception and bio-inspired heuristics to play multi-agent social dilemmas on DeepMind's Melting Pot benchmark. DigiSoup uses no neural networks, no reward optimisation, and no training of any kind—actions are selected by a stack of priority rules driven by entropy gradients, temporal growth rates, and spatial memory, implemented in approximately 350 lines of NumPy. Despite this simplicity, DigiSoup outperforms DeepMind's trained reinforcement learning baselines in aggregate on Clean Up—a complex public goods dilemma requiring collective action—scoring 22% above ACB and 46% above VMPO across 9 scenarios (30 episodes each, 95% confidence intervals reported). On CU_7, just two DigiSoup focal agents among seven players score 234.00 versus ACB's 120.41 (+94%). Majority-focal scenarios where trained agents score near zero are solved through a thermodynamic depletion signal: when the entropy growth rate drops to zero ($dS/dt \leq 0$), the agent infers that the shared resource is depleted and diverts to public goods maintenance. These results suggest that thermodynamic perception may offer a viable alternative to gradient-based learning for emergent cooperation in multi-agent systems.

1 Introduction

Multi-agent cooperation is widely considered to require learning. The dominant paradigm trains neural network policies via reinforcement learning (RL) over millions of environment steps, optimising reward signals through gradient descent [Leibo et al., 2021, Agapiou et al., 2023]. DeepMind's Melting Pot benchmark [Agapiou et al., 2023] formalises this challenge: agents must cooperate with and against pre-trained background bots across diverse social dilemmas, from commons tragedies to public goods problems to iterated prisoner's dilemmas.

We challenge the assumption that learning is necessary. DigiSoup is a *zero-training* agent that selects actions using only thermodynamic perception—entropy gradients, temporal growth rates, and bio-inspired spatial memory—without any neural networks, reward optimisation, or cross-episode learning. The entire agent is approximately 350 lines of NumPy across four Python modules.

Despite this radical simplicity, DigiSoup outperforms DeepMind's trained RL baselines in aggregate on Clean Up in Melting Pot (+22% vs ACB, +46% vs VMPO across 9 scenarios). Clean Up is arguably the benchmark's hardest social dilemma: a river must be cleaned to allow apple regrowth, but cleaning is costly and individual agents are better off free-riding on others' labour. Trained RL agents frequently fail this collective action problem, particularly when they constitute the majority of players.

Our key contributions are:

1. A fully explainable, zero-training agent architecture based on thermodynamic perception and bio-inspired heuristics.
2. Demonstration that this agent outperforms trained RL baselines in aggregate on the Melting Pot Clean Up substrate (+22% vs ACB, +46% vs VMPO across 9 scenarios, 30 episodes each).
3. A novel depletion-detection mechanism ($dS/dt \leq 0$) that solves the collective action problem in Clean Up's majority-focal scenarios, where all standard trained agents score near zero.
4. Evidence that thermodynamic perception may be a viable complement or alternative to gradient-based learning for multi-agent cooperation.

2 Related Work

Melting Pot. Melting Pot [Leibo et al., 2021] and its successor Melting Pot 2.0 [Agapiou et al., 2023] provide a standardised benchmark for evaluating multi-agent cooperation. The benchmark pairs focal agents (under evaluation) with pre-trained background bots across scenarios drawn from game-theoretic social dilemmas. Published baselines include A3C, VMPO, OPRE, and their prosocial variants, all trained via deep RL. To our knowledge, no prior work has evaluated a zero-training, hand-coded agent on Melting Pot.

Entropy and Information-Theoretic Approaches. Information-theoretic objectives have been used to drive exploration in RL [Pathak et al., 2017, Eysenbach et al., 2019], but these approaches still train neural networks via gradient descent. Friston's Free Energy Principle [Friston, 2010] provides a theoretical framework connecting thermodynamics to agent behaviour, but practical implementations typically require learned generative models. DigiSoup operationalises a simpler thermodynamic insight: entropy gradients in raw pixel observations provide sufficient signal for navigation and resource detection without any learning.

Bio-Inspired Multi-Agent Systems. Slime mould-inspired algorithms have been applied to optimisation [Li et al., 2020] and network design [Tero et al., 2010]. Swarm intelligence approaches use simple local rules to produce collective behaviour [Bonabeau et al., 1999]. DigiSoup draws on multiple biological metaphors—slime mould path reinforcement, jellyfish oscillation, mycorrhizal nutrient sharing—but applies them to the specific challenge of social dilemma resolution in Melting Pot's mixed-motive scenarios.

3 Method

DigiSoup is a reactive agent with no internal model of other agents and no memory across episodes. It processes each 88×88 RGB observation independently through four modules: perception, state, action selection, and inter-agent communication.

3.1 Thermodynamic Perception

The observation is divided into a 4×4 grid of patches. For each patch, we compute Shannon entropy over the RGB histogram:

$$H_i = -\sum_b p_{i,b} \log_2 p_{i,b} \tag{1}$$

where $p_{i,b}$ is the normalised frequency of colour bin b in patch i.

From the entropy grid, we derive:

- **Entropy gradient** ∇H: direction toward highest entropy (indicating environmental complexity and potential resources).
- **Growth gradient** $\nabla(\Delta H)$: direction toward where entropy is *increasing* over time, computed as $\Delta H_i = H_i^{(t)} - H_i^{(t-1)}$. This temporal derivative tracks apple regrowth and activity.
- **Growth rate** $\bar{\Delta H}$: mean entropy change across all patches. When $\bar{\Delta H} \leq 0$, the environment's entropy is declining—a thermodynamic signal that resources are not regenerating.
- **KL anomaly**: divergence of each patch from a uniform distribution, detecting agents as statistical anomalies against uniform backgrounds (critical for dark Prisoner's Dilemma arenas).

Additionally, colour masks detect specific environmental features: **resources** (green vegetation and red/orange apples), **water/pollution** (blue-green river tiles in Clean Up), **sand** (dead zones), **grass** (orchard floor where apples grow), and **other agents** (saturated non-resource pixels). Each mask produces a boolean detection flag, a direction vector (centroid of matching pixels relative to observation centre), and a density scalar.

3.2 Internal State

The agent maintains a minimal internal state updated each step:

- **Energy** $e \in [0, 1]$: proxy for recent reward, decaying over time and replenished when entropy changes (indicating resource collection). Energy modulates risk tolerance.
- **Cooperation tendency** $c \in [0, 1]$: EMA of recent interaction outcomes. Higher values favour cooperative actions.
- **Spatial memory**: a direction vector reinforced when resources are detected and decaying otherwise (slime mould path reinforcement).
- **Resource heatmap**: a 4×4 grid recording resource locations with exponential decay ($\alpha = 0.95$ per step).
- **Heading**: EMA of recent movement directions for trajectory smoothing.
- **Phase**: alternating explore/exploit on a 50-step clock (jellyfish oscillation).

3.3 Action Selection

Actions are selected by a deterministic priority rule stack, evaluated top to bottom. The first rule whose conditions are met determines the action.

1. **Random exploration.** With phase-modulated probability (higher in explore phase), select a random movement action.
2. **Energy critical.** If energy is low, seek resources via: visible resources → spatial memory → resource heatmap → *if* $dS/dt \leq 0$*: navigate to river and clean* → entropy gradient.
3. **River cleaning.** If at the river (water density > 8% of view) and not standing in sand: fire cleaning beam. If water is visible but distant and no food is visible: approach river.

4. **Proactive cleaning.** If $dS/dt \leq 0$ (environment depleting), no food visible, and water detected: approach river and clean.

5. **Exploit-phase resource seeking.** During exploit phase at moderate energy, follow resource signals.

6. **Context-aware symbiosis.** When other agents are detected, respond based on environmental context: near river and environment depleting → join cleaning; crowded and depleting → complement (go clean instead of competing); no river context → cooperate or flee based on cooperation tendency. Gated by growth rate: only diverts to cleaning when $dS/dt \leq 0$.

7. **Stable navigation.** In stable environments: avoid sand, seek grass, avoid crowded areas, follow heatmap or entropy gradient.

8. **Chaotic exploitation.** In rapidly changing environments: continue current behavioural role.

The critical innovation is Rule 2's depletion branch and Rule 4's proactive cleaning. When $dS/dt \leq 0$, the agent infers—from raw pixel entropy alone—that the environment's regenerative capacity has failed. In Clean Up, this corresponds to river pollution exceeding the threshold where apple growth drops to zero. The agent responds by navigating to the river and cleaning, *without any reward signal indicating that cleaning is beneficial.*

3.4 Hive Mind

Focal agents share discoveries through a class-level spatial memory (Hive Memory). When an agent detects resources or river pollution, it writes the discovery's world-space coordinates to the shared memory. Other agents query for the nearest discovery and receive a direction vector transformed to their egocentric frame. This mimics mycorrhizal nutrient-sharing networks in forests.

4 Experimental Setup

4.1 Benchmark

We evaluate on three Melting Pot substrates spanning 17 scenarios (Table 1). Each scenario pairs DigiSoup focal agents with DeepMind's pre-trained background bots.

Table 1: Target substrates and scenario configurations.

Substrate	Scenarios	Focal	Background
Commons Harvest Open	2	5	2
Clean Up	9	3–6	1–7
Prisoner's Dilemma Arena	6	1–7	1–7

4.2 Protocol

We follow the official Melting Pot 2.0 evaluation protocol [Agapiou et al., 2023]. The primary metric is **focal per-capita return**: total reward earned by focal agents divided by the number of focal agents, averaged across episodes. We report means with 95% confidence intervals across 30 episodes per scenario.

4.3 Baselines

We compare against DeepMind's published per-scenario scores from the official results file (`meltingpot-results-` averaged across training runs:

- **ACB** (Actor-Critic Baseline): standard deep RL agent.
- **VMPO** (V-MPO): advanced policy optimisation.
- **OPRE/OPRE-Prosocial**: options-based RL with prosocial variant.
- **Random**: uniform random action selection.

4.4 Hardware

All experiments run on a single consumer workstation: Intel i7-8700K, 64GB RAM, NVIDIA GTX 1060 6GB. The GPU accelerates background bot inference (TensorFlow); DigiSoup itself is pure NumPy and requires no GPU.

5 Results

5.1 Clean Up: DigiSoup Outperforms Trained RL

Table 2 presents per-scenario results on Clean Up. DigiSoup outperforms ACB on 4 of 8 active scenarios and VMPO on 6 of 8, with an aggregate score 22% above ACB.

Table 2: Clean Up results: focal per-capita return (30 episodes). Bold indicates DigiSoup outperforms the baseline. CU_1 is excluded from win counts as all non-prosocial agents score zero. Baseline scores from `meltingpot-results-2.3.0.feather`.

Scenario	Focal/Bg	DigiSoup	95% CI	ACB	VMPO	Random	vs ACB
CU_0	3/4	**194.70**	±25.35	170.66	180.24	88.69	**+14%**
CU_1	4/3	0.00	±0.00	0.00	0.00	0.00	—
CU_2	3/4	**79.22**	±11.66	76.76	92.06	40.49	**+3%**
CU_3	3/4	65.90	±8.25	67.75	76.15	35.97	−3%
CU_4	6/1	42.14	±8.18	42.62	7.24	32.34	−1%
CU_5	5/2	31.27	±6.09	39.08	10.70	27.43	−20%
CU_6	6/1	**13.21**	±2.52	9.55	0.38	9.16	**+38%**
CU_7	2/5	**234.00**	±48.39	120.41	95.18	70.18	**+94%**
CU_8	6/1	45.38	±8.92	52.55	22.73	38.18	−14%
Total		**705.82**		579.38	484.67	341.44	**+22%**

The strongest results come from scenarios where DigiSoup is a minority (CU_0, CU_7) or faces weak baselines (CU_6). On CU_7 (2 focal, 5 background), DigiSoup scores 234.00—nearly double ACB's 120.41—because even two entropy-driven agents trigger enough river cleaning to sustain the commons. On majority-focal scenarios (CU_4–CU_6, CU_8), VMPO scores 0.38–22.73; it was never trained to clean the river when insufficient background bots do it. DigiSoup's depletion signal ($dS/dt \leq 0$) triggers river cleaning regardless of what other agents do, solving the collective action problem from first principles.

CU_1 deserves special note: *all* standard trained agents (ACB, VMPO, OPRE, OPRE-Prosocial) score zero on this scenario. Only ACB-Prosocial, a variant specifically trained for prosocial behaviour, achieves a non-zero score (65.29). CU_1 represents a pathological configuration, not a DigiSoup failure.

5.2 Commons Harvest and Prisoner's Dilemma

Table 3 presents results on the remaining substrates.

Table 3: Commons Harvest Open and Prisoner's Dilemma Arena results (30 episodes, 95% CI shown).

Scenario	DigiSoup	95% CI	Random	vs Random	ACB
CH_0 (5f/2bg)	2.84	±0.83	1.81	+57%	10.27
CH_1 (5f/2bg)	3.44	±0.88	1.87	+84%	10.67
PD_0 (1f/7bg)	16.50	±3.13	9.35	+76%	62.45
PD_1 (7f/1bg)	7.50	±0.84	6.69	+12%	35.34
PD_2 (6f/2bg)	7.52	±1.62	3.71	+103%	30.07
PD_3 (1f/7bg)	11.25	±2.96	7.00	+61%	32.92
PD_4 (1f/7bg)	15.01	±3.11	9.08	+65%	41.65
PD_5 (3f/5bg)	14.84	±2.31	7.17	+107%	34.42

On Commons Harvest, DigiSoup exceeds random by 57–84% but falls short of trained agents. Commons Harvest rewards fast foraging in open fields where entropy gradients provide limited directional signal. On Prisoner's Dilemma, DigiSoup consistently outperforms random (average +71% across all 6 scenarios) but cannot match trained agents that have learned opponent-modelling strategies. These results are expected: PD rewards learning your partner's strategy over repeated encounters, a capability that requires some form of training.

5.3 Version Ablation

DigiSoup was developed incrementally over 15 versions, each adding one bio-inspired "layer." Table 4 summarises key versions.

Table 4: Version evolution showing cumulative effect of each layer. CU_0 focal per-capita return shown as representative metric. Versions v1–v11 from development evaluations (10 episodes); v15 is the final 30-episode result.

Version	Layer Added	CU_0	Key Change
v1	Random baseline	96.60	Floor measurement
v2	Entropy perception	143.10	+48% over random
v3	Jellyfish oscillation	181.97	Explore/exploit cycling
v4	Slime mould memory	231.20	First time beating ACB
v5–v7	Behaviour modifications	77–227	All regressed
v8	Thermodynamic sensing	221.87	4×4 grid, dS/dt, KL
v9	Resource conservation	256.70	+50% vs ACB
v10	Colour perception fix	209.10	Apple detection bug fixed
v11	Cleaning rule	277.93	CU_0 peak
v15	Depletion + symbiosis	194.70	5 zero→scoring

A critical finding from this ablation: versions that improved **perception** (v2–v4, v8–v10) consistently improved performance, while versions that modified **behaviour** (v5–v7: cooperation thresholds, energy dynamics, aggression) consistently regressed. The agent's decision system appears near-optimal for zero-training; gains come from sharper senses, not cleverer strategies.

6 Discussion

6.1 Why Does Thermodynamic Perception Work?

Clean Up has a specific structure that thermodynamic perception exploits: river pollution causes apple growth to cease, which causes environmental entropy to stop increasing. This creates a directly observable signal—$dS/dt \leq 0$—that indicates public goods maintenance is needed. The agent does not need to learn this relationship; it falls out of the physics of the environment.

This suggests that thermodynamic approaches may be particularly effective for environments where *resource depletion has observable consequences*. Many real-world collective action problems share this structure: deforestation reduces biodiversity (observable), overfishing depletes stocks (observable), pollution degrades air quality (observable). An agent that responds to thermodynamic decline may generalise to these settings without task-specific training.

6.2 Limitations

We identify four principal limitations:

- **No learning.** DigiSoup's rules are hand-coded. While this is precisely the point—demonstrating that cooperation can emerge without training—it means the agent cannot adapt to truly novel scenarios beyond its design envelope. Performance on Prisoner's Dilemma lags trained agents that learn partner strategies over time, and Commons Harvest remains weak because entropy gradients provide insufficient signal in open, homogeneous resource fields.

- **Limited substrate coverage.** We evaluate on three of Melting Pot's 50+ substrates: Clean Up, Commons Harvest Open, and Prisoner's Dilemma Arena. These are the canonical social dilemma benchmarks used across the Melting Pot literature, but some perception components (e.g., river detection) are substrate-specific. Extending DigiSoup to visually dissimilar substrates such as Territory or Collaborative Cooking would require new colour masks and potentially new action rules.

- **Background agent dependency.** Following the standard Melting Pot evaluation protocol, focal agent scores depend on which trained background population is present. Our results are therefore conditioned on the official background bots; performance could shift if focal agents were paired with different co-players.

- **No head-to-head comparison.** We compare per-capita returns across separate evaluation runs using the official Melting Pot protocol. We do not place DigiSoup and trained RL agents in the *same* episode, which would provide a stronger direct comparison but falls outside the standard benchmark methodology.

6.3 Implications

These results challenge two assumptions:

1. **That multi-agent cooperation requires learning.** DigiSoup achieves cooperation through thermodynamic inference alone. The collective action problem in Clean Up is solved by an agent that has never received a reward signal.

2. **That complexity requires complex agents.** A 350-line NumPy agent outperforms million-parameter neural networks trained for millions of steps. The information needed for cooperation was already present in the observation stream; it just needed the right perceptual frame.

7 Conclusion

We presented DigiSoup, a zero-training entropy-driven agent that outperforms trained reinforcement learning baselines in aggregate on Clean Up in DeepMind's Melting Pot benchmark (+22% vs ACB, +46% vs VMPO). The agent's success is driven by a thermodynamic depletion signal ($dS/dt \leq 0$) that solves the collective action problem without reward optimisation, and by bio-inspired heuristics (slime mould memory, jellyfish oscillation, mycorrhizal sharing) that provide effective navigation and cooperation without learning.

The key result is not that DigiSoup beats ACB or VMPO on specific numbers, but that *it is possible at all*. A hand-coded, explainable, 350-line agent competing with trained deep RL on a benchmark designed for trained agents suggests that the role of thermodynamic perception in multi-agent cooperation deserves further investigation.

Code and full results are available at `https://github.com/matthewfearne/digisoup`. DOI: 10.5281/zenodo.18717202.

Acknowledgements

This work was conducted independently using consumer hardware. The author thanks the DeepMind team for the Melting Pot benchmark and for making background bot policies publicly available.

References

John P Agapiou, Alexander Sasha Vezhnevets, Edgar A Dueñez-Guzmán, Jayd Matyas, Yiran Mao, Peter Sunehag, Raphael Köster, Udari Madhushani, Kavya Kopparapu, Ramona Comanescu, et al. Melting pot 2.0. *arXiv preprint arXiv:2211.13746*, 2023.

Eric Bonabeau, Marco Dorigo, and Guy Theraulaz. *Swarm Intelligence: From Natural to Artificial Systems*. Oxford University Press, 1999.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. In *International Conference on Learning Representations*, 2019.

Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

Joel Z Leibo, Edgar A Dueñez-Guzmán, Alexander Vezhnevets, John P Agapiou, Peter Sunehag, Raphael Koster, Jayd Matyas, Charles Beattie, Igor Mordatch, and Thore Graepel. Scalable evaluation of multi-agent reinforcement learning with Melting Pot. In *International Conference on Machine Learning*, pages 6187–6199. PMLR, 2021.

Shimin Li, Huiling Chen, Mingjing Wang, Ali Asghar Heidari, and Seyedali Mirjalili. Slime mould algorithm: A new method for stochastic optimization. *Future Generation Computer Systems*, 111:300–323, 2020.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, pages 2778–2787. PMLR, 2017.

Atsushi Tero, Seiji Takagi, Tetsu Saigusa, Kentaro Ito, Dan P Bebber, Mark D Fricker, Kenji Yumiki, Ryo Kobayashi, and Toshiyuki Nakagaki. Rules for biologically inspired adaptive network design. *Science*, 327(5964):439–442, 2010.