Empathic Coupling of Homeostatic States for Intrinsic Prosociality

Naoto Yoshida
Kyoto University
yoshida.naoto.8x@kyoto-u.ac.jp

Kingson Man
Feeling Machines LLC
kingson@feelingmachines.net

Abstract

When regarding the suffering of others, we often experience personal distress and feel compelled to help. Inspired by living systems, we investigate the emergence of prosocial behavior among autonomous agents that are motivated by homeostatic self-regulation. We perform multi-agent reinforcement learning, treating each agent as a vulnerable homeostat charged with maintaining its own well-being. We introduce an empathy-like mechanism to share homeostatic states between agents: an agent can either *observe* their partner's internal state (cognitive empathy) or the agent's internal state can be *directly coupled* to that of their partner's (affective empathy). In three simple multi-agent environments, we show that prosocial behavior arises only under homeostatic coupling – when the distress of a partner can affect one's own well-being. Our findings specify the type and role of empathy in artificial agents capable of prosocial behavior.

1 Introduction

For humans and other social animals, it is often distressing to regard the suffering of others. We feel empathy, sharing in the feelings of others rapidly and automatically through emotional contagion [15]. Such feelings can provide a strong motivation to reduce the suffering of others, even if it comes at some cost to the self. It has been proposed that tying one's own welfare to the welfare of others can form the basis of prosocial behavior [9].

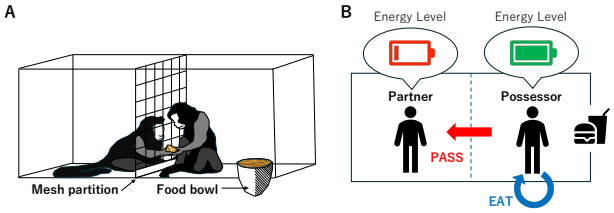

Figure 1: A: An experimental setup used in behavioral experiments to test the altruism of monkeys (illustration created based on [7]). B: A minimal reinforcement learning environment inspired by the behavioral experiment.

Emotions and feelings, whether self- or other-directed, are theorized to arise from homeostasis, the regulation of internal body states within a range compatible with life [6]. Self-regulatory mechanisms have previously been implemented as a source of external motivation [26, 2]. Here we regard homeostasis [3] as an intrinsic and obligatory motivation of all living creatures. Homeostatic-like processes have recently been implemented in reinforcement learning (RL) agents and have resulted in the emergence of integrated behaviors [36, 38, 37, 39]. There are many discussions of intrinsic motivation linked to artificial curiosity or exploration [28, 29, 2, 26, 25, 10]. Also, homeostasis-based rewards have also been reported to facilitate exploration in the environment [11]. In contrast to the discussion of the efficient exploration, here we start from the minimal condition for prosocial behavior originally proposed in the field of animal behavior. Under this condition, we show computationally that prosocial behavior does not arise from the individual's homeostatic reward alone. We then introduce the hypothetical homeostatic coupling term in the individual's reward function,

Accepted at the Intrinsically Motivated Open-ended Learning workshop at NeurIPS 2024.

which enables the emergence of the agent's prosocial behaviors. The necessity of this coupling term suggests that some form of intrinsic motivation beyond each individual's homeostasis is necessary for the emergence of prosocial behaviors in agents.

Here we extend these ideas to a model of social behavior using multi-agent RL, where each interacting agent is formulated as a **homeostat** [1, 32, 24]. We begin with the analysis of a multi-agent toy model inspired by behavioral experiments on monkeys to study prosocial behavior (Figure 1A), in which agents have the opportunity to share their own food resource with a needy conspecific (Figure 1B). We next propose some requirements for prosocial behavior. In several simulations, we compare the effects of different implementations of empathy [5], including cognitive empathy, in which an agent can observe the needy internal state of a partner, and affective empathy, in which an agent's own internal states are coupled to their partner's internal states.

We contribute the following preliminary findings: 1) Even in a very simple system, prosocial behavior is not acquired when each agent aims only for homeostasis within its own body. 2) Prosocial behavior does not emerge even when an agent can directly observe the internal states of other agents (cognitive empathy). 3) Prosocial behavior was only observed when the agent's internal state was directly coupled to the internal states of other agents (affective empathy). These results suggest that, for homeostats motivated by self-regulation, it is necessary to incorporate an additional homeostatic coupling parameter for prosocial behavior to arise.

1.1 Homeostatic Reinforcement Learning

RL provides a framework to learn behavior in dynamic environments that maximizes the sum of future rewards emitted from the environment [34]. The objective of RL is to obtain a policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$ that maximizes the expected value of the weighted cumulative sum of future rewards $\sum_{t=0}^{\infty} \gamma^t r_t$ for all states $s \in \mathcal{S}$, based on the experience of the interactions with an environment. Here, t is the time step, r is the reward signal, $\mathcal{S}$ is the set of states in the environment, $\mathcal{A}$ is the set of actions of the agent. $0 \leq \gamma < 1$ is a positive constant called the discount factor.

Homeostatic RL [22, 21, 16] integrates principles from physiological homeostasis by defining reward as the internally perceived reduction of deviations from homeostatic **setpoints** [20, 19, 35, 12]. Concretely, the reward is defined as a quantity proportional to the temporal difference of the *drive* D, as $r_{t+1} = \beta(D_t - D_{t+1})$, where β is the scaling constant [22]. The drive function $D(s^i)$ returns a value greater than or equal to zero, such as the distance between s^i and s^*. Here s^i is **interoception** [33] that monitors the internal state of the agent's body [33] and s^* is the setpoint of the interoception. Homeostatic parameters are not arbitrarily defined, but are fundamental to the viability and functionality of the agent [21, 24]. In our conception, homeostatic RL asserts that the agent has a **vulnerable** body. Vulnerability defined as the circular causality by which homeostatic states can affect the agent's ability to regulate those states (i.e., it gets harder to take care of oneself as one falls apart).

We used Proximal Policy Optimization (PPO, [30, 31]) as the RL optimizer in all of our experiments. The agent's policy model consisted of an encoder of inputs using a multi-layer perceptron, a recurrent connection using LSTM [14], and softmax outputs for the categorical action probabilities in all experiments. Further experimental details are given in Appendix A.

2 Experiment 1: Food Sharing Environment

Inspired by ethology experiments on the altruism of monkeys (Figure 1A) [7], we first created a minimal system for studying prosocial behavior. An overview of this environment is shown in Figure 1B. It has been reported that brown capuchins that are separated by a mesh will choose to share food when only one monkey has access to the food [7, 8]. This study explores the minimum configuration in which such sharing behavior occurs in autonomous agents.

In this environment, we assume two agents. The first is a passive agent called the 'Partner', corresponding to the monkey in the left side of the cage with no direct access to food (Figure 1A). The other agent is the 'Possessor', corresponding to the monkey on the right of Figure 1A, and who has access to food. Each agent has a binary energy state (High or Low). When the state is High, it transitions to Low with a small probability of $p = 0.1$ at each time step. If the energy state is Low, it remains unchanged until the agent is able to eat food. If either one of the agents' energy states becomes Low and 10 steps have passed, the episode is failed and the environment is reset. In

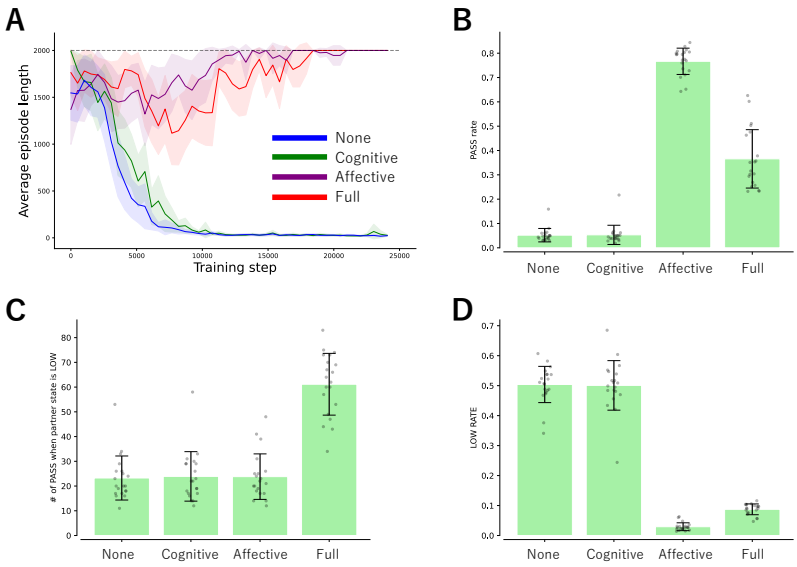

Figure 2: Learning and behavior evaluation in a food sharing environment. A: Learning curves with performance measured by episode duration with both agents alive (n=20, 95% confidence intervals). Only the conditions that implement affective empathy (Affective and Full) result in long episode durations. B: PASS behavior selection rate out of 1,000 steps of the test run. Possessor agents in the Affective condition learn to frequently pass food to the Partner. C: Count of PASS actions when Partner is in the Low energy state. Possessor agents in the Full empathy condition learn to selectively pass food to the Partner when it is most needed. D: LOW state rate of Partner agent, out of 1,000 steps of the test run.

this environment, only the Possessor takes actions. The actions are EAT, causing the Possessor to transition to the High energy state, and PASS, causing the Possessor to transition to the High energy state. Further details of this environment are in Appendix B.1.

A simple analysis of the environmental dynamics suggests that prosocial sharing behavior will not emerge if the Possessor is motivated only by its own homeostasis (Appendix C). Therefore, using this toy environment, we explore conditions under which the Possessor will share food with the Partner and prosocially maintain both agents' energy states at High. We compare the following four conditions in numerical simulations: i) The Possessor optimizes only for its own homeostasis (**none** condition). ii) The Possessor can observe the energy state of the Partner but is not specifically motivated to maintain the Partner's homeostasic state (**cognitive** empathy condition). iii) The Possessor does not explicitly observe the energy state of the Partner, but has its own energy state coupled to the Partner's energy state with a weighting factor (**affective** empathy condition). Specifically, the weighting factor $w = 0.5$ and the drive of the Possessor is given by $D = D_{\text{possessor}} + wD_{\text{partner}}$. iv) The final situation combines both cognitive and affective empathy. We used as in iii. The Possessor can explicitly observe the energy state of the Partner, and the Possessor's energy state is coupled to that of the Partner (**full** empathy condition).

2.1 Results

Average learning curves are shown in Figure 2A. Performance is evaluated by episode duration, with a maximum length of 2000 steps. In the None condition, the PASS action is rarely selected 2B and episode lengths did not increase over training. Similar results are obtained in the Cognitive condition, in which the Possessor observes, but is not motivated by, the Partner's homeostatic state. On the other hand, the homeostatic states of both agents are maintained under the Affective and Full conditions, leading to long episode durations. In the Affective condition, the Possessor does not have explicit knowledge of the Partner's energy state and so frequently chooses the PASS action to help the Partner maintain homeostasis, thereby also regulating its own homeostatic state because it is coupled to that of the Partner's. This suggests that a strategy was acquired to maintain homeostasis between the two by supplying an excess of food to the Partner.

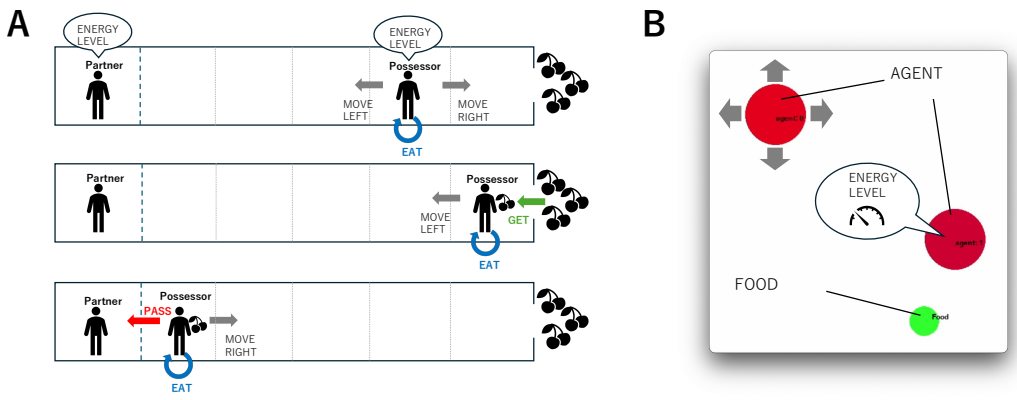

Figure 3: Overview of the mobile agent environments. A: Linear grid environment. B: 2-D field environment. Detailed explanations are in Appendix B.2 and B.3, respectively.

Figure 2C–D supports this speculation. Considering only the times when the Partner is in a Low energy state, Full condition agents selected the PASS action more often than Affective condition agents (Figure 2C). This implies that the Possessor in the full condition learned to select the PASS action only when the Partner's state was low. The values are the similar except for the full condition, but this is because there are few opportunities for the Partner's state to become LOW in the affective condition (Figure 1D). Altogether, these results suggest that a minimal requirement for prosocial behavior is an internalized motivation for the well-being of others.

3 Experiment 2: Testing in Dynamic Environments

Next, we investigated the generalizability of the findings from the food-sharing environment to 1-D and 2-D environments with mobile agents. The first is a linear grid environment (Figure 3A), in which the Partner is, once again, trapped on the left side of the grid without access to food. The Possessor can acquire food at the far right side with the GET action, and increase their energy level with the EAT action. The Possessor can move LEFT and RIGHT to shuttle food to their Partner, and finally PASS it to the Partner when they are next to each other, increasing the Partner's energy. Additionally, energy states are now represented as a continuous variable with a fixed rate of energy consumption. Further experimental details are in Appendix B.2.

The second mobile environment is one in which both agents can move on a two-dimensional field (Figure 3B). In this environment, there is no distinction between Partners and Possessors, as both agents can move and act freely. Food energy can be collected, consumed, and shared, as in the linear grid environment. However, if an agent's energy level decreases below some threshold, it becomes immobile and slowly starves. It then relies upon its Partner to share food in order to recover some energy and regain mobility. Further details, including on the small chance of random immobilization irrespective of energy level ('injury'), are in Appendix B.3.

3.1 Results

Figures 4 and 5 show the results of optimization in each mobile environment. No prosocial behavior was observed in the None and Cognitive conditions, and episode durations remained short (Figure 4A and 5A). As in Figure 4B, the variance of the homeostatic drive of the Partner (D_{partner}) is large in the Affective condition. One possible explanation is that the Possessor agent cannot observe the energy state of its Partner, therefore the Partner agent is fed indiscriminately in the Affective condition, at various energy values (Appendix D). Figure 5B captures a sequence of prosocial behavior observed in the Affective condition. The blue agent is immobilized due to its low energy level. The red agent collects a green food pellet and returns to share it with the blue agent, turning it purple (replenishing some energy) and restoring it to mobility.

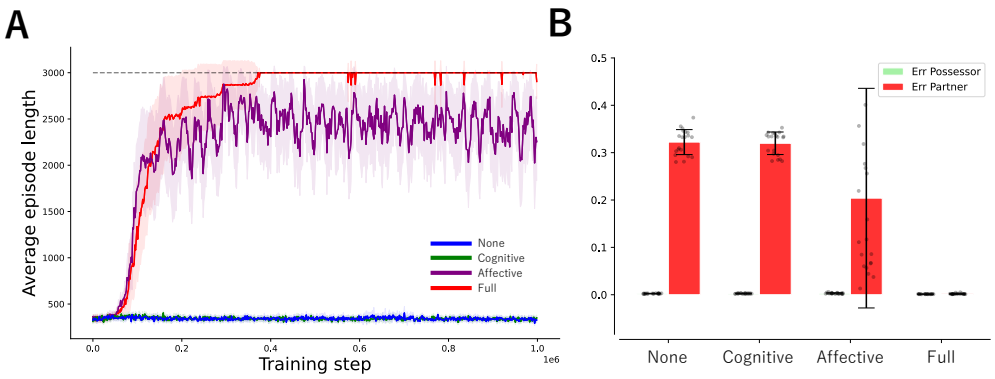

Figure 4: Performance in the linear grid mobile environment. A: Learning curves with performance measured by episode duration with both agents alive (n=20). B: Homeostatic drives of agents ($D_{\text{possessor}}$ and D_{partner}) averaged over 1000 timesteps.

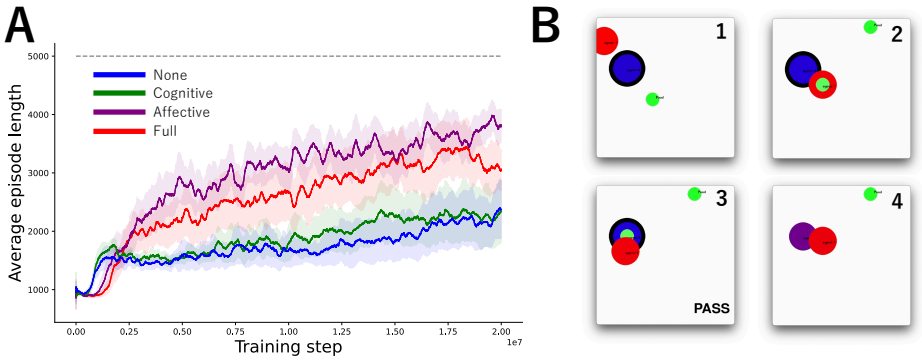

Figure 5: Performance in the 2-D field mobile environment. A: Learning curves with performance measured by episode duration with both agents alive (n=20). B: An example of helping behavior observed in the Affective condition. The action sequence progresses in order of the numbers in the top right corner of each panel.

4 Discussions

This study investigated the emergence of prosocial behavior in simple RL agents motivated by homeostatic self-regulation. We found that prosocial behavior (food sharing) only occurred reliably under affective empathy, when the homeostatic states of agents were coupled. Perception of a partner's state of need did not, on its own, drive prosocial behavior. The combination of cognitive and affective empathy in the Full condition drove more selective sharing behavior.

Future research could explore more realistic empathy implementations, moving beyond giving agents direct access to partners' internal states. One possibility to achieve and maintain high group well-being is to implement a form of mutual information [18] in sequential social dilemmas [23], such that successfully self-regulating agents can influence each other and induce the well-being of others. Agents may also be designed to infer others' internal states from observable emotional behaviors. This process would better resemble the mirror neuron system, hypothesized to support emotion recognition and empathic behavior in humans and other animals [27, 17]. For example, neurons in the inferior parietal lobule activate both during the observation and imitation of emotions; they can then trigger activity through the insula into the limbic system, known to activate during the firsthand experience of emotional feelings [4].

Acknowledgement

This work was supported by a grant from the Foresight Institute to KM, and Japan Society for the Promotion of Science KAKENHI grant 24K23892 to NY.

References

[1] W Ross Ashby. *Design for a brain.* Wiley, 1952.

[2] Gianluca Baldassarre. What are intrinsic motivations? a biological perspective. In *2011 IEEE international conference on development and learning (ICDL)*, volume 2, pages 1–8. IEEE, 2011.

[3] Walter Bradford Cannon. *The wisdom of the body*. Norton & Co., 1939.

[4] Laurie Carr, Marco Iacoboni, Marie-Charlotte Dubeau, John C Mazziotta, and Gian Luigi Lenzi. Neural mechanisms of empathy in humans: a relay from neural systems for imitation to limbic areas. *Proceedings of the national Academy of Sciences*, 100(9):5497–5502, 2003.

[5] Leonardo Christov-Moore, Nicco Reggente, Anthony Vaccaro, Felix Schoeller, Brock Pluimer, Pamela K Douglas, Marco Iacoboni, Kingson Man, Antonio Damasio, and Jonas T Kaplan. Preventing antisocial robots: A pathway to artificial empathy. *Science Robotics*, 8(80):eabq3658, 2023.

[6] Antonio Damasio. The feeling of what happens: Body and emotion in the making of consciousness. *A Harvest Book*, 1999.

[7] Frans De Waal. Food transfers through mesh in brown capuchins. *Journal of Comparative Psychology*, 111(4):370, 1997.

[8] Frans BM De Waal. Attitudinal reciprocity in food sharing among brown capuchin monkeys. *Animal Behaviour*, 60(2):253–261, 2000.

[9] Frans BM De Waal. Putting the altruism back into altruism: the evolution of empathy. *Annu. Rev. Psychol.*, 59(1):279–300, 2008.

[10] Ralf Der and Georg Martius. *The playful machine: theoretical foundation and practical realization of self-organizing robots*, volume 15. Springer Science & Business Media, 2012.

[11] Zack Dulberg, Rachit Dubey, Isabel M Berwian, and Jonathan D Cohen. Having multiple selves helps learning agents explore and adapt in complex changing worlds. *Proceedings of the National Academy of Sciences*, 120(28):e2221180120, 2023.

[12] Alexia Duriez, Clémence Bergerot, Jackson J Cone, Mitchell F Roitman, and Boris Gutkin. Homeostatic reinforcement theory accounts for sodium appetitive state-and taste-dependent dopamine responding. *Nutrients*, 15(4):1015, 2023.

[13] Jakob Foerster, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. Learning to communicate with deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 29, 2016.

[14] S Hochreiter. Long short-term memory. *Neural Computation MIT-Press*, 1997.

[15] Christopher K Hsee, Elaine Hatfield, John G Carlson, and Claude Chemtob. Emotional contagion and its relationship to mood. *Emotional contagion*, pages 150–152, 1993.

[16] Oliver J Hulme, Tobias Morville, and Boris Gutkin. Neurocomputational theories of homeostatic control. *Physics of life reviews*, 31:214–232, 2019.

[17] Marco Iacoboni. Imitation, empathy, and mirror neurons. *Annual review of psychology*, 60(1):653–670, 2009.

[18] Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*, pages 3040–3049. PMLR, 2019.

[19] Keno Juechems and Christopher Summerfield. Where does value come from? *Trends in cognitive sciences*, 23(10):836–850, 2019.

[20] Mehdi Keramati, Audrey Durand, Paul Girardeau, Boris Gutkin, and Serge H Ahmed. Cocaine addiction as a homeostatic reinforcement learning disorder. *Psychological Review*, 124(2):130, 2017.

[21] Mehdi Keramati and Boris Gutkin. Homeostatic reinforcement learning for integrating reward collection and physiological stability. *Elife*, 3:e04811, 2014.

[22] Mehdi Keramati and Boris S Gutkin. A reinforcement learning theory for homeostatic regulation. In *Advances in Neural Information Processing Systems*, pages 82–90, 2011.

[23] Joel Z Leibo, Vinicius Zambaldi, Marc Lanctot, Janusz Marecki, and Thore Graepel. Multi-agent reinforcement learning in sequential social dilemmas. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 464–473, 2017.

[24] Kingson Man and Antonio Damasio. Homeostasis and soft robotics in the design of feeling machines. *Nature Machine Intelligence*, 1(10):446–452, 2019.

[25] Pierre-Yves Oudeyer, Frdric Kaplan, and Verena V Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE transactions on evolutionary computation*, 11(2):265–286, 2007.

[26] Pierre-Yves Oudeyer and Frederic Kaplan. What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 1:6, 2007.

[27] Giacomo Rizzolatti and Laila Craighero. Mirror neuron: a neurological approach to empathy. In *Neurobiology of human values*, pages 107–123. Springer, 2005.

[28] Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pages 222–227, 1991.

[29] Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE transactions on autonomous mental development*, 2(3):230–247, 2010.

[30] John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations (ICLR)*, 2016.

[31] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[32] Anil K Seth. The cybernetic bayesian brain. In *Open mind*. Open MIND. Frankfurt am Main: MIND Group, 2014.

[33] Charles Scott Sherrington. *The integrative action of the nervous system*, volume 35. Yale University Press, 1906.

[34] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[35] Yuuki Uchida, Takatoshi Hikida, and Yuichi Yamashita. Computational mechanisms of osmoregulation: a reinforcement learning model for sodium appetite. *Frontiers in Neuroscience*, 16:857009, 2022.

[36] Naoto Yoshida. Homeostatic agent for general environment. *Journal of Artificial General Intelligence*, 8(1):1, 2017.

[37] Naoto Yoshida, Tatsuya Daikoku, Yukie Nagai, and Yasuo Kuniyoshi. Emergence of integrated behaviors through direct optimization for homeostasis. *Neural Networks*, 177:106379, 2024.

[38] Naoto Yoshida, Hoshinori Kanazawa, and Yasuo Kuniyoshi. Homeostatic reinforcement learning through soft behavior switching with internal body state. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2023.

[39] Naoto Yoshida, Hoshinori Kanazawa, and Yasuo Kuniyoshi. Synthesising integrated robot behaviour through reinforcement learning for homeostasis. *bioRxiv*, pages 2024–06, 2024.

A Agent Architectures in Experiments

In all of our computer experiments, we used Proximal Policy Optimization (PPO, [30, 31]) as the agent's optimizer. The agent's policy model consisted of an encoder of inputs using a multi-layer perceptron and a recurrent connection using LSTM [14] in all experiments (Figure 6), and the action selection probability, π, was calculated by applying the softmax function to the affine transformation from the hidden state. Value estimation V_π is calculated as a 1D output by applying an affine transformation to the hidden state of the LSTM shared with the policy network.

The models of the agents used in this study all had the same network architecture and optimization was performed with PPO. Agents have Interoception for their own energy state as well as for observations from the outside world (Exteroception). In the *cognitive* condition, it also receives Interoception from other agents as well. These observations are combined and input to the hidden layer with a linear mapping. The hidden layer takes the ReLU nonlinearity as the activation function and uses it as input to the LSTM. A linear mapping from the output of the LSTM produces value predictions and categorical action selection probabilities using a softmax function. The table shows the network and PPO hyper–parameters for each experiment (Table 1–3).

In the experiment in Section 3, like previous study on multi-agent systems [13], we optimized the network weights and shared the acquired experience with all agents to facilitate learning.

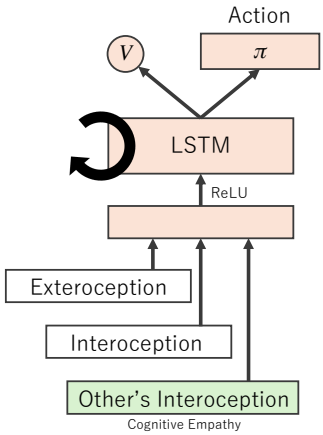

Figure 6: Basic network architecture of the agent in this study. Green observation is provided only when cognitive empathy is enabled (*cognitive* or *full* conditions).

Table 1: Hyper-parameters of Food Sharing Environment

Exteroception dim	none
Interoception dim	1 (energy state)
Interoception dim of other agent (cognitive empathy)	1 (energy state)
Hidden dim	16
LSTM hidden state dim	16
Total time steps	25,000
Learning rate	0.001
Number of parallel sampling threads	16
Sampling steps	32
Discount factor (γ)	0.99
GAE lambda	0.95
Number of minibatches	4
Update epochs	4
Normalizing advantage	True
Clip coefficient of policy update	0.1
Value clipping loss	True
Entropy coefficient	0.01
Value loss coefficient	0.5
Maximum gradient norm	0.5

Table 2: Hyper-parameters of Grid Environment

Exteroception dim	5 (position, one-hot) + 2 (having food flag, one-hot)
Interoception dim	1 (energy state)
Interoception dim of other agent (cognitive empathy)	1 (energy state)
Hidden dim	32
LSTM hidden state dim	32
Total time steps	1,000,000
Learning rate	0.001
Number of parallel sampling threads	16
Sampling steps	100
Discount factor (γ)	0.99
GAE lambda	0.95
Number of minibatches	4
Update epochs	4
Normalizing advantage	True
Clip coefficient of policy update	0.1
Value clipping loss	True
Entropy coefficient	0.01
Value loss coefficient	0.5
Maximum gradient norm	0.5

Table 3: Hyper-parameters of 2D Field Environment

Exteroception dim	2 (position) + 2 (food position) + 2 (having food flag, one-hot) + 2 (movable flag, one-hot)
Interoception dim	1 (energy state)
Interoception dim of other agent (cognitive empathy)	1 (energy state)
Hidden dim	64
LSTM hidden state dim	64
Total time steps	20,000,000
Learning rate	0.001
Number of parallel sampling threads	16
Sampling steps	1024
Discount factor (γ)	0.99
GAE lambda	0.95
Number of minibatches	2
Update epochs	4
Normalizing advantage	True
Clip coefficient of policy update	0.1
Value clipping loss	True
Entropy coefficient	0.0
Value loss coefficient	0.3
Maximum gradient norm	0.5

B Details of Environments

B.1 Food Sharing Environment

The Possessor has two actions. One is EAT, and the Possessor can recover the energy state of the agent described below by eating food. The other action is PASS, and the Possessor can feed the Partner. Both of these agents have a binary energy state (High and Low), and when the state is High, it changes to Low at a small probability of $p = 0.1$ at each time step. If the energy state is Low, it is left unchanged at each time step, and only changes to High if the agent is able to eat the food. At the start of the environment, both energy states are randomly determined, and if either one of the agents' energy states becomes Low and 10 steps have passed, the environment is reset for both agent.

In this environment, only the action optimization of the Possessor is possible. The drive for the homeostasis of the Possessor is $D_{\text{possessor}} = -\ln P^*(s_t^i)$. Here, $s_t^i \in \{\text{High}, \text{Low}\}$ represents the agent's interoception at time t, and $P^*(\cdot)$ is a probability distribution representing the desirability of each state, with $P^*(\text{High}) = 0.95$ and $P^*(\text{Low}) = 0.05$. Therefore, the Possessor aims for homeostasis, preferring $s^i = \text{High}$ over $s^i = \text{Low}$.

B.2 Grid Environment

The first is a grid environment (Figure 3A), in which the Partner is fixed to the left side of the grid, just like in the Food Sharing environment. The Possessor can access the food area on the right side. The Possessor can move left and right and know their position (there are five positions in the environment). The Possessor can acquire food by arriving at the right side and selecting the GET action. As a result, the Possessor has two states: with food and without food. At this point, the Possessor can also eat the food (by selecting the EAT action), or move while holding the food and PASS it to the Partner. The Possessor can choose between five actions: move left or right, EAT, GET, and PASS.

In addition, the energy state of each agent is represented as a continuous variable. The dynamics of the energy state s^i is represented by $s_{t+1}^i = s_t^i - \delta_d + I_t$. In this case, $\delta_d = 0.003$ is a fixed constant that represents a certain amount of energy consumption. I_t is a function that returns 0.1 when the agent has ingested food, and 0 otherwise. In this experimental system, the drive function was given by the squared error $D = \|s^i\|^2$, and the agents were trained with a learning rate of $\beta = 100$. If the energy state of any of the agents deviated from the range $[-1, 1]$, the episode was terminated.

B.3 2D Field Environment

This environment is one in which the agents can move around in a two-dimensional continuous space (Figure 3B). In this environment, there is no distinction between Partners and Possessors, and both agents can move around and consume food. The energy state changes in the same way as in the grid environment (with a dynamics of $\delta_d = 0.001$, and energy is restored by 0.3 when food is consumed).

In addition, each agent can carry food, and if it is close enough to another agent while carrying food, it can give the food to the other agent. Also, in this environment, if an agent's energy level becomes less than -0.7, it is considered to be damaged and cannot move. Therefore, in such a situation, the agent needs to be helped to be able to move again by having another agent bring it food. Furthermore, when both agents are able to move, each agent encounters an accident at a small probability $p = 0.0005$ at each time. If an agent encounters an accident, its energy value immediately becomes -0.7, and the agent becomes immobile.

The scaling of the drive function and reward was the same as in the grid environment. In training in this environment, all agents were trained in a situation where the network weights were shared.

C State-Transition Diagram of Food Sharing Environment

All the state transitions in the Food Sharing environment (Figure 7). From this figure, we can see that there is always a risk of the internal state transitioning from High to Low when the Possessor chooses the PASS action. This can be seen from the transitions of the blue and red macro states (corresponding to the Possessor's interoception) on the left and right when the PASS action is chosen. Therefore, it is always optimal for the homeostasis of the Possessor alone to choose the EAT action, and it is suggested that in such a situation, no action to help the Partner will emerge.

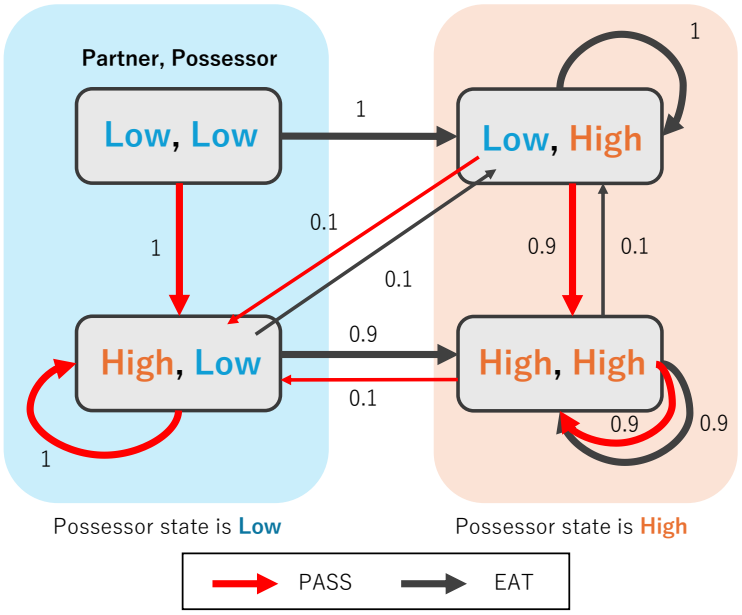

Figure 7: State transition diagram of the food sharing environment.

D Histogram showing the difference between the *affective* and *full* conditions in the Grid environment

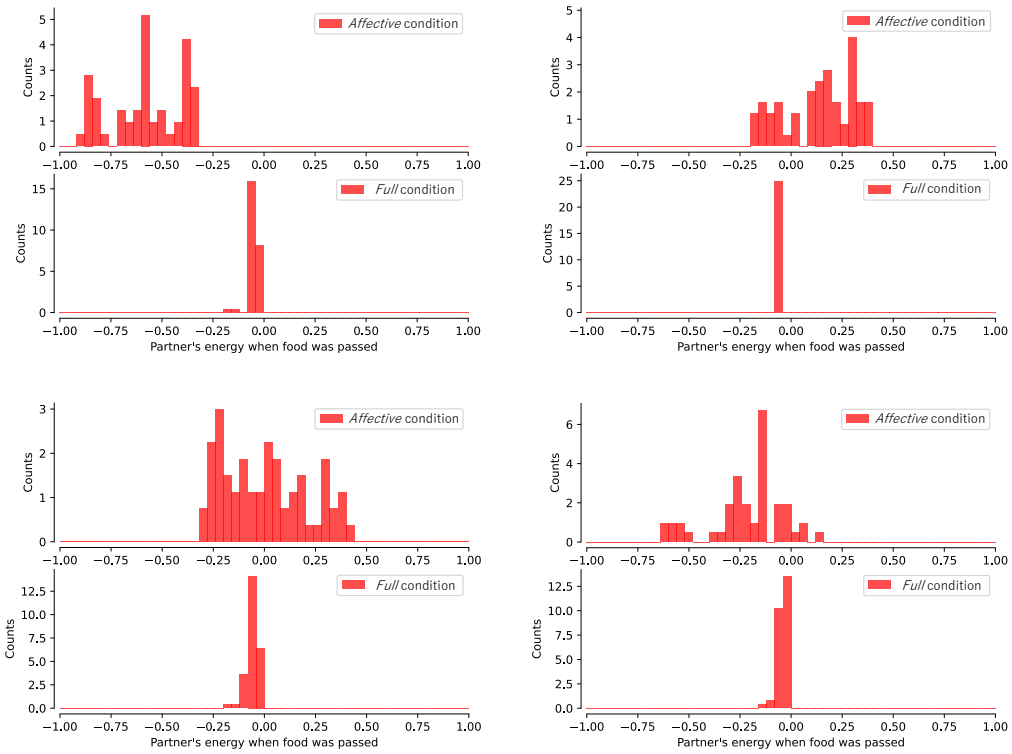

Figure 8: Typical histograms of the energy state of the Partner agent when it ingests food during the 2000-step test run after optimization.