EMBODIMENT PERSPECTIVE OF REWARD DEFINITION FOR BEHAVIOURAL HOMEOSTASIS

Naoto Yoshida[1,2], **Tatsuya Daikoku**[2], **Yukie Nagai**[2,3], **and Yasuo Kuniyoshi**[1]
[1]Department of Mechano-Informatics, The University of Tokyo
[2]International Research Center for Neurointelligence (WPI-IRCN), The University of Tokyo
[3]Institute for AI and Beyond, The University of Tokyo
{n-yoshida, kuniyosh}@isi.imi.i.u-tokyo.ac.jp
{daikoku.tatsuya, nagai.yukie}@mail.u-tokyo.ac.jp

ABSTRACT

In this work, we propose a *neural homeostat*, a neural machine that stabilises the internal physiological state through interactions with the environment. Based on this framework, we demonstrate that behavioural homeostasis with low-level continuous motor control emerges from an embodied agent using only rewards computed by the agent's local information. Using the bodily state of the embodied agent as the reward source, the complexity of the reward definition is 'outsourced' into the coupled dynamics of the bodily state and the environment. Therefore, our definition of the reward is simple, but the optimised behaviour of the agent can be surprisingly complex. Our contributions are 1) an extension of homeostatic reinforcement learning to enable continuous motor control using deep reinforcement learning; 2) a comparison of homeostatic reward definitions from previous studies, where we found that homeostatic rewards using the difference of the drive function performed best; and 3) a demonstration of the emergence of adaptive behaviour from low-level motor control through direct optimisation of the homeostatic objective.

1 INTRODUCTION

The definition of rewards for general-purpose autonomous agents has been a long-debated problem (Lewis et al., 2010; Baldassarre, 2011; Silver et al., 2021). Moreover, defining rewards for desired purposes is a known problem in the reinforcement learning (RL) community (Clark & Amodei, 2016; Amodei et al., 2016), and complex reward definitions are necessary for sophisticated tasks. In this work, we introduce the embodiment perspective of the reward definition. Our definition of the reward is simple, but the optimised behaviour of the agent is surprisingly complex. We show that the complexity of the reward definition can be outsourced into the coupled-dynamics of the bodily state and the environment. Indeed, we simulated the internal dynamics of the agent as an additional complexity in our experiments. Nevertheless, there is no need to actually perform those simulations in embodied agents because such dynamics are inherent in real autonomous agents (like robots, animals).

Homeostasis, the stabilisation of the internal body state, is considered a fundamental function of animals. Richter (1943) suggested adaptive behaviour in animals as the realisation of homeostasis in living organisms. In addition, Hull (1943) proposed in his classical study that the reduction of 'drive' defines the learning motivation, and the drive arises due to the physiological needs of the body. Because of the simplicity of the idea, homeostasis as a foundation of adaptive behaviours has been adopted by many researchers in ethology (Barnard, 2004; McFarland & Bösser, 1993), nutritional science (Simpson et al., 2010), human-agent interaction (Blumberg, 1997; Ogata & Sugano, 2000b; Breazeal, 2002) and computational neuroscience (Keramati & Gutkin, 2011; Gu & FitzGerald, 2014; Seth, 2014; Keramati & Gutkin, 2014; Pezzulo et al., 2015; Hulme et al., 2019; Man & Damasio, 2019). An intuitive explanation of the emergence of adaptive behaviour in homeostasis is shown in Figure 1.

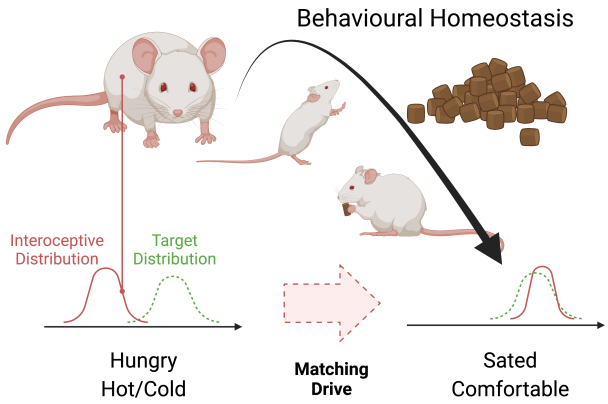

Figure 1: Our hypothetical mechanism of the emergence of the adaptive motor control from the homeostatic objective. The agent has a target signal distribution (setpoint) of specific sensor stimuli (green-dotted distribution). The agent receives the current sensor stimuli from the inside of the body (red distribution). Then, he tries to match the current sensation to the target. In order to do so, the agent has to control the inside of the body directly (homeostasis) or interact with the environment through behaviours (behavioural homeostasis, allostasis). (Figure created with BioRender.com)

In spite of numerous consistent proposals of homeostasis as an elementary objective, the computational approach to the direct optimisation of this objective is limited (Konidaris & Barto, 2006; Keramati & Gutkin, 2014). This limitation is because the optimisation for behavioural homeostasis naturally needs to treat multimodal sensations. These sensations include interoceptive signals (visceral signals, energy levels, and water levels) in addition to proprioception (joint states) and exteroception (vision, sounds). In addition, interoceptive signals can be regarded as context variables in multitask RL (Hallak et al., 2015; Sodhani et al., 2021). In our homeostatic RL setting (Keramati & Gutkin, 2014), the context variable changes continuously through agent-environment interactions. Furthermore, the behavioural objective (drinking water, eating fruit for carbohydrates, cooling the body temperature) dynamically changes depending on the body's internal state (water level, nutritional state, body temperature). Due to these differences in dynamics, RL for homeostasis differs from single-task multimodal RL and formal multitask RL.

These fundamental complexities of the homeostatic problem require the radical simplification of the value function estimation or small-scale experiments, as determined in previous studies (Whitehead et al., 1993; Bersini, 1994; Konidaris & Barto, 2006; Keramati & Gutkin, 2014). Recent advances in deep RL approaches for continuous motor control have enabled researchers to use the universal approximation function, which has led to impressive results in robotics (Schulman et al., 2016), the control of the anatomical rodent model (Merel et al., 2020), and the humanoid control with visual image inputs from low-level continuous motor signals (Merel et al., 2019). However, the application of deep RL to homeostatic control problems has only recently been applied to discrete action domains (Yoshida, 2017).

In this study, our methodological focus was on scaling up homeostatic RL to the high-dimensional motor control domain. We hypothesise that complex adaptive motor control emerges from embodied homeostatic reward signals. We demonstrated our method in two novel environments. The first environment is an extended version of the classical homeostatic environment *two-resource problem* (TRP) (Spier, 1997). The second is a temperature regulation environment in which the agent needs to regulate body temperature in addition to energy homeostasis. Our contributions are 1) an extension of homeostatic reinforcement learning to enable continuous motor control using deep RL; 2) a comparison of homeostatic reward definitions from previous studies, where we found that homeostatic rewards using the difference of the drive function performed best; and 3) a demonstration of the emergence of adaptive behaviour from low-level motor control through direct optimisation of the homeostatic objective. Our neural realisation of a homeostatic system is reminiscent of the *homeostat* (Ashby, 1952) developed in a classical cybernetic study. Therefore, we refer to our class of neural system as *neural homeostat*.

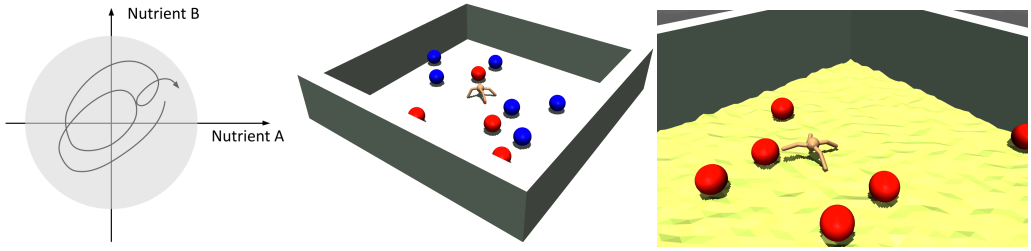

Figure 2: Two-resource problem environment. (left) A sketch of the internal physiological state dynamics of the agent. If the internal state is inside the grey area, the agent is 'alive'. If it reaches the outside of this area, the agent 'dies', and the episode terminates. (middle) Overview of the field of the two-resource problem environment. There are randomly distributed food resources (four red balls and six blue balls) and a quadruped robot in the arena. (right) The body temperature regulation environment. In this environment, the agent needs to regulate both the energy resource and the core body temperature.

2 Neural Homeostat for Behavioural Homeostasis

In this section, we introduce a class of architecture called a *neural homeostat*. This architecture describes the minimal architecture required to realise behavioural homeostasis with the continuous motor control as an integration of homeostasis and deep RL. An overview of our problem setting is presented in Figure 2. An agent is in an environment (middle and right panels) and has an internal physiological state in the body (left panel). The agent's goal lies inside the agent's body: stabilisation of the internal state in a particular area.

The agent receives three types of multidimensional observations: exteroception x^e, a proprioception x^p and an interoception x^i[1]. The agent constructs a reward function using these observations, and interoception is used to construct the fundamental reward function r_{homeo} for homeostasis.

2.1 Homeostatic Reward with Reward Shaping

RL is a machine learning framework that maximises the expectation of the agent's future cumulative sum of rewards $\sum_{t=0}^{\infty} \gamma^t r_r$ through interactions with the environment (Sutton & Barto, 2018). Here, t is the time step, r_t is the reward, and $0 \leq \gamma < 1$ is the discount factor. An agent receives an observation x from the environment in a single time step and returns an action u from the parametric stochastic policy π_θ. RL algorithms optimise the policy through interactions with the environment.

Homeostatic RL intends to stabilise the internal state of the agent through RL (Keramati & Gutkin, 2014; Hulme et al., 2019). To achieve this goal, we employ interoception as a source of the elementary reward for homeostasis. In our study, we employed the quadratic homeostatic drive $D(x_t^i) = ||x_t^i - x_*^i||^2$, where x_*^i is an interoceptive target. We used the differential form of the homeostatic reward function, which has also been proposed in previous studies (Keramati & Gutkin, 2014; Hulme et al., 2019):

$$r_{\text{homeo}} = \beta_1 (D(x_t^i) - D(x_{t+1}^i)), \tag{1}$$

where $\beta_1 > 0$ is a positive constant. We call this reward definition the homeostatic reward with reward shaping, or simply the 'homeostatic-shaped' reward. The quadratic form of the drive $r = -D(x_t^i)$ is the reward definition proposed by other studies (McFarland & Houston, 1981; McFarland & Bösser, 1993), and its probabilistic interpretation is also provided in another study (Yoshida, 2017). In Supplementary Material D, we explain that the differential form of the reward definition can be derived by using policy-invariant transformations (Ng et al., 1999) from the quadratic reward. Furthermore, as a representation of the prior distribution for motor outputs and agent posture, we applied proprioceptive cost in addition to homeostatic reward:

$$r_{\text{cost}} = -\beta_2 ||\tilde{x}_t^p - \tilde{x}_*^p||^2 - \beta_3 ||u_t||^2, \tag{2}$$

[1]Minimally exteroception and proprioception may be treated equally in our framework.

where $\tilde{x}^p_*$ and $\tilde{x}^p$ are the default (torso in upright position) and current posture, u is the agent's action, and β_2 and β_3 are positive constants. We believe that these posture information can be computed within the agent, and that there is no need for an oracle-like mechanism outside the robot (Lee et al., 2020). The total reward function is described as:

$$r_{\text{total}} = r_{\text{homeo}} + r_{\text{cost}}. \tag{3}$$

We used $\beta_2 = 0.005$, and $\beta_3 = 0.001$ throughout the study.

We can show that our problem setting can be explained using divergence minimisation (Hafner et al., 2020b) between the actual distribution and the Gaussian target distribution. In this context, the cost terms for interoceptive and proprioceptive correspond to Gaussian target distributions for those inputs. The details of the derivation are provided in Supplementary Material D.

2.2 Alternative Reward Definitions

Homeostatic control can be achieved using several alternative definitions. To confirm the effectiveness of the homeostatic-shaped reward definition, it was compared with three alternative reward definitions. A straightforward example is the quadratic homeostatic reward without reward shaping $r_{\text{homeo}} = -\beta_1 D(x^i_t)$, which is treated in the previous studies (McFarland & Houston, 1981; Keramati & Gutkin, 2014; Yoshida, 2017). We will call this reward definition the (fundamental) 'homeostatic' reward. The second alternative is a quadratic reward with a baseline shift. In this definition, we employ a reward baseline $b > 0$, and the quadratic reward is shifted by $\tilde{r}_{\text{homeo}} = -\beta_1 D(x^i_t) + b$. Because the quadratic term cannot be positive, this reward definition generates positive- and negative-reward areas in the interoception space. We note that this baseline shift does not change the optimal policy. A similar definition of this reward function was also employed in a previous study for homeostatic control (Bersini, 1994) in a simple grid environment, and it performed the best among the definitions proposed in that environment. This definition will henceforth be denoted as the 'homeostatic-biased' reward definition. Finally, the third definition is the 'Cart-Pole' style reward (Sutton & Barto, 2018). The reward is zero except for at the terminal state (death, resetting state), which provides $-\beta_1$. We denote this as 'cart-pole' reward definition. The cart-pole is the simplest definition among others, and this definition may have a theoretical advantage in that the terminal reward bounds the value function. However, this definition does not utilise the information of the internal state, except for the terminal states.

All four definitions of the reward function are task-general regarding the arbitrariness of their solutions for homeostatic control. In the experiment section, we compare the reward definitions to compare their learning performance in terms of survival time steps in the environment. In preliminary experiments, we conducted a hyperparameter search of the scaling of the homeostatic term of the reward β_1 and the bias term b in $\{0.01, 0.1, 1, 10, 100, 1000\}$, and then we compared the results with the best parameter settings.

2.3 Agent Architectures and Optimization

In our experiments, we employed the proximal policy optimisation (PPO) with the generalised advantage estimator (GAE) (Schulman et al., 2017; 2016). We used a fully connected architecture with two hidden layers for both the policy network π_θ and the value prediction network V_ϕ, which have 256 and 64 units in their hidden layers, with hyperbolic activation units, respectively. A beta distribution $\text{Beta}(\alpha_\theta, \beta_\theta)$ is used as the output of the policy network (Chou et al., 2017; Hsu et al., 2020). $\alpha_\theta(x)$ and $\beta_\theta(x)$ are branched outputs of the policy network, which are parameterized by θ with an observation x. Because the output of the beta policy is restricted in the d-dimensional space $[0, 1]^d$, outputs are scaled into $[-1, 1]^d$ as actions that are used in the environment.

The objective to be maximized in the update is described as

$$J(\theta, \phi) = \hat{\mathbf{E}}_{\pi_{\text{old}}} \left[L^{CLIP}(\theta) - c_1 L^{VF}(\phi) + c_2 S(\pi_\theta) - c_3 \tilde{D}(\pi_{\text{old}} || \pi_\theta) \right], \tag{4}$$

where θ is the policy parameter and ϕ is the value prediction parameter. $\hat{\mathbf{E}}_{\pi_{\text{old}}}[\cdot]$ represents the empirical average sampled using the previous policy π_{old}. $L^{CLIP}(\theta)$ is the surrogate loss of PPO for the policy improvement, $L^{VF}(\phi)$ is the value-prediction loss, and $S(\pi_\theta)$ is the entropy bonus.

$\tilde{D}(\pi_{\text{old}}||\pi_\theta)$ is the approximated Kullback-Leibler divergence penalty between current and the previous policy, which is known to stabilise the optimisation (Hsu et al., 2020). Detailed descriptions of the architecture and objectives are provided in Supplementary Material A. We use the same hyperparameters $c_1 = 0.5$, $c_2 = 0.001$ and $c_3 = 0.001$ throughout this study.

We used the Adam optimizer with epsilon parameter 10^{-5} for optimisation with a learning rate that started from 3×10^{-4} and linearly decreased to 10^{-5} along with 500 PPO iterations. 3×10^5 training batch data were collected using ten worker threads, and a mini-batch size of 5×10^4 was used for the stochastic gradient descent. Additional hyperparameters are provided in Supplementary Material A.

3 EXPERIMENTS

Two experiments were conducted. The first was a two-resource foraging experiment that extended the classical homeostatic problem into a continuous motor control domain. The second experiment was a thermal regulation experiment in which the agents were required to regulate their core body temperature while regulating energy homeostasis.

3.1 TWO-RESOURCE PROBLEM WITH CONTINUOUS MOTOR CONTROL

This section introduces TRP (Figure 2, middle), which is proposed in the context of the theoretical ethology (Spier, 1997). In this environment, there is an agent and randomly distributed food resources in the field. The agent can act in the field, which was developed using a dynamics simulator (Todorov et al., 2012). The agent model is based on the quadruped robot 'Ant' (Schulman et al., 2016) from the food-gathering environment (Duan et al., 2016; Li et al., 2020). We used the 'low-gear' version of the Ant asset, in which the stability of the motion is improved. In our experiments, the action of the agent was the motor torque of each joint. The dimension of the control d was eight, and the control space was normalised so that the control is in the d-dimensional unit cube $u \in [-1, 1]^d$.

The agent has a two-dimensional continuous internal state that corresponds to the nutritional state of the agent. Two types of food resources (four red balls and six blue balls in the field) correspond to the dimensions of the nutritional state of the agent. Unequal numbers of red and blue balls require the agent to balance the food collection instead of simply collecting the food types equally. The agent consumed the nutrient resources linearly with time steps. If the agent's body sphere gets close enough to a food resource, this food is consumed, and the nutritional state is recharged with a predefined quantity. New food resources were randomly generated. Our experiment assumes that the agent directly observes the internal state as the interoceptive sensory input x^i.

We utilised a simplified metabolic model of the internal nutritional state described in a previous study (Konidaris & Barto, 2006). The updates of the nutritional state x^i for both of the red and blue resources are described as

$$x^i_{t+1} = x^i_t - \delta^i_{\text{default}} + \delta^i_{\text{food}} I^i_t, \tag{5}$$

where $\delta^i_{\text{default}}$ is the default consumption of nutritions, and δ^i_{food} is the inlet of the nutrition when the agent captures the food resource. I_t is one if the agent gets close enough (less than 1m in the simulator); otherwise zero. We used the same parameters used in a previous study (Konidaris & Barto, 2006); $\delta^i_{\text{default}} = 0.00015$ and $\delta^i_{\text{food}} = 0.1$. A single episode starts from nutritional states uniformly sampled from $\mathcal{U}[-\frac{1}{6}, \frac{1}{6}]$ for each nutrient. An episode terminates if any one of the internal variables exceeds the viable range $[-1, 1]$, and the environment is reset afterwards.

The agent's observations are composed of a 40-dimensional exteroception x^e, a 27-dimensional proprioception x^p and a two-dimensional interoception x^i. Exteroception is the agent's perceptual signals outside of the body, composed of range sensor stimuli (20 different directions around the agent) for two kinds of food resources. Proprioception is the observation that reports the self-movement and positions of the agent's body, and we assume that this involves the agent's joint angles, rotational and positional speed, the height and posture information of the torso. Interoception is the direct observation of the internal nutritional state of the agent, which is composed of a two-dimensional continuous vector.

3.2 BODY TEMPERATURE REGULATION WHILE FORAGING

In this experiment, we introduced an alternative homeostatic setting inspired by the thermoregulation of animals. The agent needs to regulate the core body temperature through interactions in the environment while maintaining energy homeostasis. Here, we only implemented the simplified dynamics of the thermal system to demonstrate the proof-of-concept of our method. The actual biological thermoregulation system is sophisticated and complex (Porter et al., 1973; Terrien et al., 2011; Tan & Knight, 2018). The environment is shown in the right panel of Figure 2. Food resources are randomly distributed over a square arena with randomly generated terrain. In the thermal regulation experiment, we used temperature dynamics in addition to one-dimensional nutritional dynamics.

We implemented a dynamic model of core body temperature in the environment, inspired by thermodynamic models of animals in thermal biology (Porter et al., 1973; Fei et al., 2012) and thermal models of electric motors (Venkataraman et al., 2005). We model the dynamics of the body temperature τ as

$$C\frac{d\tau}{dt} = \delta Q(\tau, u, u_{\mathrm{ev}}), \tag{6}$$

where τ is the animal's core body temperature and C is the heat capacity of the body. δQ is the amount of heat that is added to the body of the agent.

The agent has a nine-dimensional action. u is the eight-dimensional motor output, and $u_{\mathrm{ev}} \in [-1, 1]$ is a one-dimensional 'evaporative' action that controls the heat dissipation rate. A detailed description of the calculation of δQ is provided in Supplementary Material C. The core body temperature is normalised to the map $[307, 315]$ in Kelvin degrees to $[-1, 1]$. Thus, the setpoint of the normalised body core temperature is zero, which corresponds to 38 °C (311 K). The agent has a one-dimensional nutrient state, which has the same dynamics as the previous experiment. The interoceptive signal x^i is is defined as a two-dimensional signal composed of a one-dimensional nutrient state and the normalised core temperature of the agent. The agent receives the same exteroception and proprioception, in which the range finder stimulus represents the distance to food resources in this environment (red balls, 20 dimensions).

4 RESULTS

4.1 HOMEOSTATIC FORAGING WITH CONTINUOUS MOTOR CONTROL IN THE TRP

The performance summaries of each reward setting in the TRP environment are shown in Figure 3 (top). Because reward definitions vary, we compared the performance using the average length of the episodes during the 10 test runs. In addition, we manually terminated test runs if the agent exceeded 60,000 steps because trained agents can survive an arbitrarily long time in the environment, and sometimes the evaluation process takes an unreasonably long time.

From Figure 3 (top-left), we can observe that the homeostatic-shaped, homeostatic, and homeostatic-biased reward could improve the episode length. In addition, the homeostatic-shaped setting achieved the best performance. However, the cart-pole definition failed to improve the policy. Because the cart-pole definition cannot provide intermediate reinforcing signals when food resources are captured, we suspect that this definition of homeostatic reward can be effective in small-scale environments. Figure 3 (top-right) shows the growth of the total number of environments resetting during the 500 PPO iterations with ten parallel agents. This criterion can capture the property that sampler agents can also survive in the environment (Shimoguchi & Kurashige, 2019). We can observe that the growth in the number gradually decreases with the training iteration with successful reward definitions.

The bottom panels of Figure 3 show the sampled behaviour of the agent after 500 PPO iterations with the homeostatic-shaped reward. As shown in the plot (bottom-right panel), internal states fluctuate within a specific range, and the agent starts to capture food resources requested according to the level of the nutrient state at that time step. In an additional experiment, we found that the simple food-capturing reward (+1 reward if the agent receives any food resources, otherwise zero) could not result in homeostatic behaviour (Supplementary Material E).

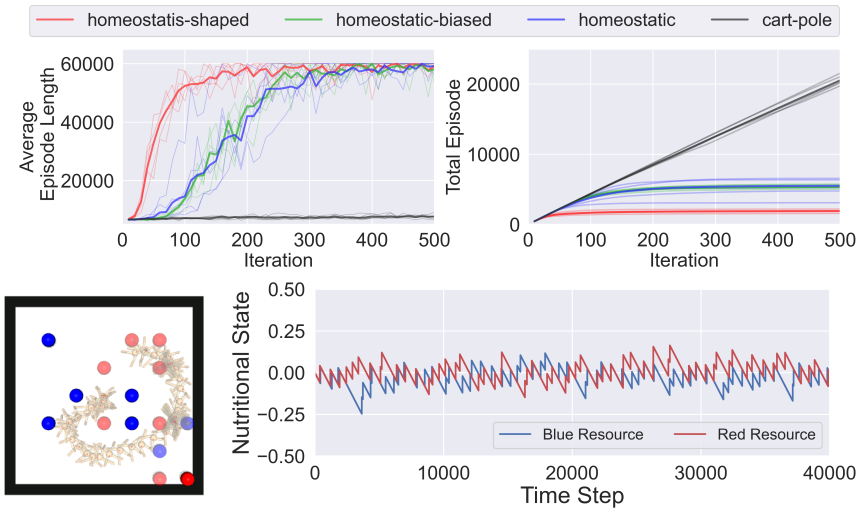

Figure 3: Training performance of the methods. The horizontal axis is the number of PPO iterations. Vertical axes reflect the average episode lengths (top left) and increases in the number of episode terminations (top right; lower is better). Thin lines represent the results of five independent runs, and thick lines represent mean performance. The bottom panels are examples of the behaviour of the agent after the optimization. (bottom left) An example of the foraging behaviour of the agent. (bottom right) Internal dynamics of internal states throughout 40,000 decision steps.

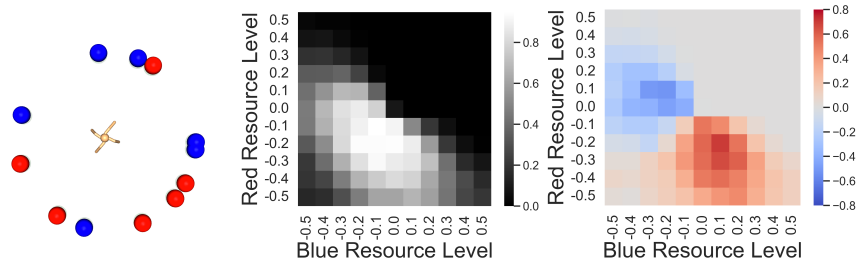

Figure 4: Behavioural experiment of the trained agent in the TRP. In this experiment, the agent's interoception was clamped at specific values, and the behavioural preference toward food resources was observed. (left) The overview of the initial condition of the experiment. The positions of red and blue balls were randomly changed between trials. The result was averaged over five individually trained agents. (middle) Food collection tendency of the agent, depending on the specific clamped interoception. (right) Interoception-dependent preference of the food collection of the agent. The blue and red areas represent the behavioural preference of the agents toward corresponding food resources.

4.1.1 Behavioural Preference in Response to the Interoception

To verify that the optimised agent selects food resources depending on the agent's interoception instead of simply collecting resources randomly, we manually clamped the agent's nutritional state at specific levels and observed the choice of food resources. We found that agents were still active under this condition, and they changed their behaviours depending on their interoception. The setting of this experiment is shown in Fig.4 (left). An agent was located in the centre of the field, and six red and blue resources were randomly scattered around the agent at a fixed distance. This process did not include the training process, and all the agent parameters were fixed during the experiment. An additional explanation is provided in Supplementary Material F.

The middle panel in Figure 4 shows the possibility that the agent captured any one of the red or blue resources. We can observe that the food capture diminishes in the range of $x^i_{\mathrm{red}} > 0$ & $x^i_{\mathrm{blue}} > 0$, which explain the suppression of the food capturing behaviour when nutrient states are both 'ful-

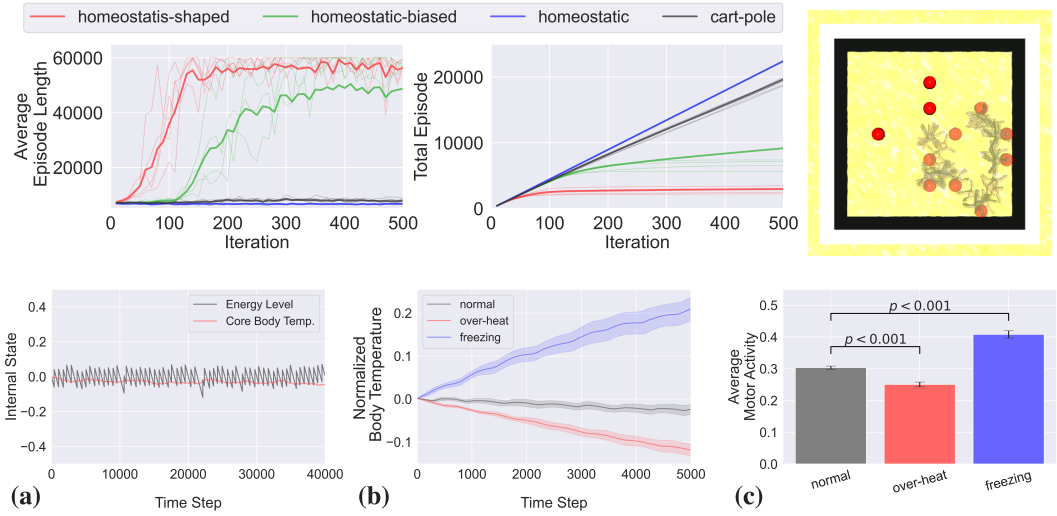

Figure 5: Training performance of the methods. (top-left) Average episode length of 5 runs. (top-middle) Growth of total episodes along with PPO iterations. (top-right) An example of the time-lapse of the agent's behaviour when trained under the homeostatic-shaped condition. a) Changes to the energy level (grey) and the core body temperature (red) of the agent (trained under the homeostatic-shaped condition) during 40,000 time steps. b) Changes to normalized body temperatures by time step. The grey curve represents the normal condition (no clamping). The red and blue curves represent the results of the observation clamping of the body temperature at 0.2 ('over-heat') and -0.2 ('freezing'). c) Motor activities of the normal (grey), 'over-heat' (red), and 'freezing' (blue) conditions with temperature-clamped observations. Motor activities were calculated by the root mean squared sum of the motor actions.

filled'. In addition, the panel shows that the agent stops foraging when both the red and blue resource deficits are large. We have provided an additional discussion on this point in Supplemental Material F. The right panel shows that the agent's preference for food resources depends on interoception. The red area indicates a preference for red resources, which is the same for the blue area. This panel clearly demonstrates the agent's interoception-dependent strategy; the agent takes the appropriate food when any one of the nutrient levels become negative. To observe the interplay between the agent's behavioural strategy and the environmental condition, we tried the same experiment using agents trained in the TRP with the (red: 5, blue: 5) condition. In this preliminary trial, we observed a similar results with (red: 4, blue: 6) condition (Supplementary Material G). Further investigation is needed to observe the effect of resource conditions in the environment toward the behavioural strategy of the agent.

4.2 Body Temperature Control while Foraging

The left two panels at the top of Figure 5 show the performance improvement results that occurred in the experiment. We observed that the agents with homeostatic-shaped and homeostatic-biased settings successfully controlled both their energy level and core body temperature. However, as shown in the panels, the homeostatic and cart-pole settings could not achieve survival behaviours in this environment. The top-right and bottom-left panels show an example of the behaviour of the agent trained with the homeostatic-shaped condition.

To demonstrate that the agent successfully obtained thermal homeostasis, we clamped the agent's body temperature interoception to specific fixed values. Then, we observed changes in the actual body temperature along with the time steps. We note that these clamping conditions did not significantly change the energy homeostasis, as observed in the TRP experiments. Panel (b) represents the 10-average results of the clamped agents ('over-heat': $\tau = 0.2$, 'freezing': $\tau = -0.2$) and the normal agent (grey). As expected, the agent with the 'over-heat' observation (red) constantly tried to decrease the body temperature, and the agent with the 'freezing' observation (blue) kept

increasing the body temperature. Panel (c) shows the average motor activities of each condition. We can see that the motor activities of the freezing-observation agent (blue) was significantly large, and the agent with an overheated observation (red) moved less than the normal agent. In addition to this experiment, we observed that the agent regulated the body temperature in response to external perturbations of body temperature (Supplementary Material H). These results would demonstrate that an agent actively regulates body temperature using motor control. Finally, we compared the behaviour of the evaporative action, but we could not find consistent behaviours between the trained agents in this setting. We suspect that the contribution of evaporative actions was limited in this experimental condition and the control of motor activities was sufficient for the homeostasis, as the environmental temperature function supports cooling of the agent's body temperature.

5 Discussions and Related Studies

Computational neuroscientists have explained homeostasis as a predictive error reduction process regarding the body's internal state (Gu & FitzGerald, 2014; Seth, 2014; Keramati & Gutkin, 2014; Pezzulo et al., 2015; Hulme et al., 2019). For example, Stephan connected the allostatic mechanism (Sterling, 1988; 2012) with an active inference process (Friston et al., 2017) and conducted a simple low-dimensional computational experiment of interoceptive control by using hierarchical generative models (Penny & Stephan, 2014; Stephan et al., 2016). Although homeostatic RL aims at a similar perspective, it more directly utilises RL theories to explain homeostatic behaviours (Hulme et al., 2019). In our study, with the help of the advance of deep RL, we conducted a direct optimisation of the behavioural homeostasis with continuous motor control to scale up the homeostatic RL.

The RL approach to the regulation of the internal states of an agent was pioneered by Bersini (1994). The divergence minimisation perspective of behaviour optimisation has been discussed in the machine learning community (Hafner et al., 2020b; Ghasemipour et al., 2020). The concept of matching the target distribution and the actual distribution is also linked with active inference (Friston et al., 2009; Friston & Ao, 2012) and our perspective (Supplementary Material D). From this perspective, the application of divergence minimisation to interoception is new in neural homeostat. Furthermore, this naturally results in a generic objective for autonomous agents. Ogata & Sugano (1997; 2000a) reported the emergence of autonomous heat control using cooling fans in a robot from the viewpoint of homeostatic control in robotics. Our thermal regulation experiment can be regarded as a generalisation of such research.

Finally, from the perspective of the reward definition problem in RL, our homeostatic reward definition r_{homeo} is simple. Indeed the introduction of the dynamics of the body's internal states may appear to be a technical definition of reward after all. However, these dynamics are realised by the physical dynamics of the embodied agent (batteries, metabolic system, heat capacity, etc.). Therefore there is no need to implement them in the real world. Rather, the complexity of the definition in the reward function is 'outsourced' into the coupled dynamics of the agent's body and the environment.

6 Conclusion

In this research, using a deep RL approach, we scaled up the idea of homeostatic RL as a principle of reward definition for an autonomous embodied agent. Our neural homeostat demonstrated the realisation proof of homeostatic behaviours from low-level motor control, which can emerge from the interoceptive reward definition by coupling the dynamics of the agent's internal state and the environment. Our environments require only minimal cognitive capabilities, and reactive agents are sufficient to solve foraging tasks. The extension of agents to treat more advanced cognitive functions should be addressed in future research.

Acknowledgments

This research was supported by JST CREST 'Cognitive Mirroring' (Grant Number: JPMJCR16E2), Institute for AI and Beyond, The University of Tokyo, World Premier International Research Centre Initiative (WPI), MEXT, JSPS KAKENHI Grant Number 20K22676, Japan, and The University of Tokyo Toyota-Dwango Scholarship for Advanced AI Talents. The funding sources had no role in the decision to publish or prepare the manuscript.

REFERENCES

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*, 2016.

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

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

Christopher J Barnard. *Animal behaviour: mechanism, development, function and evolution*. Pearson Education, 2004.

Hugues Bersini. Reinforcement learning for homeostatic endogenous variables. *From animals to animats*, 3:325–333, 1994.

Bruce Mitchell Blumberg. *Old tricks, new dogs: ethology and interactive creatures*. PhD thesis, Massachusetts Institute of Technology, 1997.

Cynthia L Breazeal. *Designing sociable robots*. MIT press, 2002.

Po-Wei Chou, Daniel Maturana, and Sebastian Scherer. Improving stochastic policy gradients in continuous control with deep reinforcement learning using the beta distribution. In *International conference on machine learning*, pp. 834–843. PMLR, 2017.

Jack Clark and Dario Amodei. Faulty reward functions in the wild, 2016. URL `https://openai.com/blog/faulty-reward-functions`.

Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International conference on machine learning*, pp. 1329–1338. PMLR, 2016.

Teng Fei, Andrew K Skidmore, Valentijn Venus, Tiejun Wang, Martin Schlerf, Bert Toxopeus, Sjef Van Overjijk, Meng Bian, and Yaolin Liu. A body temperature model for lizards as estimated from the thermal environment. *Journal of Thermal Biology*, 37(1):56–64, 2012.

Karl Friston and Ping Ao. Free energy, value, and attractors. *Computational and mathematical methods in medicine*, 2012, 2012.

Karl Friston, Thomas FitzGerald, Francesco Rigoli, Philipp Schwartenbeck, and Giovanni Pezzulo. Active inference: a process theory. *Neural computation*, 29(1):1–49, 2017.

Karl J Friston, Jean Daunizeau, and Stefan J Kiebel. Reinforcement learning or active inference? *PloS one*, 4(7):e6421, 2009.

Yasuhiro Fujita, Prabhat Nagarajan, Toshiki Kataoka, and Takahiro Ishikawa. Chainerrl: A deep reinforcement learning library. *Journal of Machine Learning Research*, 22(77):1–14, 2021.

Seyed Kamyar Seyed Ghasemipour, Richard Zemel, and Shixiang Gu. A divergence minimization perspective on imitation learning methods. In *Conference on Robot Learning*, pp. 1259–1277. PMLR, 2020.

Xiaosi Gu and Thomas HB FitzGerald. Interoceptive inference: homeostasis and decision-making. *Trends in Cognitive Sciences*, 18(6):269–70, 2014.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, pp. 2451–2463, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations (ICLR)*, 2020a.

Danijar Hafner, Pedro A Ortega, Jimmy Ba, Thomas Parr, Karl Friston, and Nicolas Heess. Action and perception as divergence minimization. In *NeurIPS Deep RL Workshop 2020*, 2020b.

Assaf Hallak, Dotan Di Castro, and Shie Mannor. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259*, 2015.

Chloe Ching-Yun Hsu, Celestine Mendler-Dünner, and Moritz Hardt. Revisiting design choices in proximal policy optimization. In *Workshop on Real World Challenges in RL (RWRL@NeurIPS)*, 2020.

Clark L Hull. *Principles of Behavior: An Introduction to Behavior Theory*. New York: Appleton-Century-Crofts, 1943.

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

Liyiming Ke, Sanjiban Choudhury, Matt Barnes, Wen Sun, Gilwoo Lee, and Siddhartha Srinivasa. Imitation learning as f-divergence minimization. In *International Workshop on the Algorithmic Foundations of Robotics*, pp. 313–329. Springer, 2020.

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

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

George Konidaris and Andrew Barto. An adaptive robot motivational system. In *From Animals to Animats 9*, pp. 346–356. Springer, 2006.

Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning quadrupedal locomotion over challenging terrain. *Science robotics*, 5(47), 2020.

Richard L Lewis, Satinder Singh, and Andrew G Barto. Where do rewards come from? In *Proceedings of the International Symposium on AI-Inspired Biology*, pp. 2601–2606, 2010.

Alexander Li, Carlos Florensa, Ignasi Clavera, and Pieter Abbeel. Sub-policy adaptation for hierarchical reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2020.

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

David McFarland and Tom Bösser. *Intelligent behavior in animals and robots*. MIT Press, 1993.

David McFarland and Alasdair Houston. *Quantitative ethology*. Pitman Advanced Pub. Program, 1981.

Josh Merel, Arun Ahuja, Vu Pham, Saran Tunyasuvunakool, Siqi Liu, Dhruva Tirumala, Nicolas Heess, and Greg Wayne. Hierarchical visuomotor control of humanoids. In *International Conference on Learning Representations (ICLR)*, 2019.

Josh Merel, Diego Aldarondo, Jesse Marshall, Yuval Tassa, Greg Wayne, and Bence Ölveczky. Deep neuroethology of a virtual rodent. In *International Conference on Learning Representations (ICLR)*, 2020.

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

Tetsuya Ogata and Shigeki Sugano. Emotional behavior adjustment system in robots. In *Proceedings 6th IEEE International Workshop on Robot and Human Communication. RO-MAN'97 SENDAI*, pp. 352–357. IEEE, 1997.

Tetsuya Ogata and Shigeki Sugano. The adaptive motion by the endocrine system model in an autonomous robot. In *International Symposium on Adaptive Motion of Animals and Machines*, number E30, August 2000a.

Tetsuya Ogata and Shigeki Sugano. Emotional communication robot: Wamoeba-2r emotion model and evaluation experiments. In *Proceedings of the International Conference on Humanoid Robots*. Citeseer, 2000b.

Will Penny and Klaas Stephan. A dynamic bayesian model of homeostatic control. In *International Conference on Adaptive and Intelligent Systems*, pp. 60–69. Springer, 2014.

Giovanni Pezzulo, Francesco Rigoli, and Karl Friston. Active inference, homeostatic regulation and adaptive behavioural control. *Progress in neurobiology*, 134:17–35, 2015.

W. P. Porter, J. W. Mitchell, W. A. Beckman, and C. B. DeWitt. Behavioral implications of mechanistic ecology. *Oecologia*, 13(1):1–54, 1973. doi: 10.1007/BF00379617.

Curt P Richter. Total self-regulatory functions in animals and human beings. *Harvey Lecture Series*, 38(63):1942–1943, 1943.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In Y. Bengio and Y. LeCun (eds.), *International Conference on Learning Representations (ICLR)*, 2014.

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.

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

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

Yuya Shimoguchi and Kentarou Kurashige. Decision making on robot with multi-task using deep reinforcement learning for each task. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, pp. 3460–3465. IEEE, 2019.

David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. Reward is enough. *Artificial Intelligence*, pp. 103535, 2021.

Stephen J Simpson, David Raubenheimer, Michael A Charleston, Fiona J Clissold, et al. Modelling nutritional interactions: from individuals to communities. *Trends in Ecology & Evolution*, 25(1): 53–60, 2010.

Shagun Sodhani, Amy Zhang, and Joelle Pineau. Multi-task reinforcement learning with context-based representations. In *the 38th International Conference on Machine Learning (ICML)*, 2021.

Emmet Spier. *From Reactive Behaviour to Adaptive Behaviour: Motivational Models for Behaviour in Animals and Robots*. PhD thesis, Balliol College, University of Oxford, 1997.

Klaas E Stephan, Zina M Manjaly, Christoph D Mathys, Lilian AE Weber, Saee Paliwal, Tim Gard, Marc Tittgemeyer, Stephen M Fleming, Helene Haker, Anil K Seth, et al. Allostatic self-efficacy: a metacognitive theory of dyshomeostasis-induced fatigue and depression. *Frontiers in human neuroscience*, 10:550, 2016.

Peter Sterling. Allostasis: a new paradigm to explain arousal pathology. *Handbook of life stress, cognition and health*, 1988.

Peter Sterling. Allostasis: a model of predictive regulation. *Physiology & behavior*, 106(1):5–15, 2012.

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

Chan Lek Tan and Zachary A Knight. Regulation of body temperature by the nervous system. *Neuron*, 98(1):31–48, 2018.

Jeremy Terrien, Martine Perret, and Fabienne Aujard. Behavioral thermoregulation in mammals: a review. *Front Biosci*, 16(4):1428–1444, 2011.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5026–5033. IEEE, 2012.

Booma Venkataraman, Bruce F. Godsey, William Premerlani, Eugene Shulman, Manish Thakur, and Rene Midence. Fundamentals of a motor thermal model and its applications in motor protection. In *58th Annual Conference for Protective Relay Engineers, 2005.*, pp. 127–144. IEEE, 2005.

Steven Whitehead, Jonas Karlsson, and Josh Tenenberg. Learning multiple goal behavior via task decomposition and dynamic policy merging. In *Robot learning*, pp. 45–78. Springer, 1993.

Eric Wiewiora. Potential-based shaping and Q-value initialization are equivalent. *J. Artif. Intell. Res.(JAIR)*, 19:205–208, 2003.

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