Supplementary Material

A Optimization Details and Hyperparameters

Proximal policy optimisation (PPO) with the generalised advantage estimator (GAE) (Schulman et al., 2017; 2016) was used in our experiments. In comparison experiments, identical hyperparameters were used, except for the reward scaling. The network architecture is a fully connected architecture with two hidden layers for both the policy network π_θ and the value prediction network V_ϕ (Figure S1). Networks have 256 and 64 units in hidden layers with hyperbolic activation units ($\tanh$). All weight parameters are initialised by orthogonal initialisation (Saxe et al., 2014), and biases are initialized by zero. A beta distribution $\mathrm{Beta}(\alpha_\theta, \beta_\theta)$ is employed as the output of the policy network (Chou et al., 2017; Hsu et al., 2020). $\alpha_\theta(x)$ and $\beta_\theta(x)$ are the branched outputs of the policy network with an observation x after the second hidden layer. Following Chou et al. (2017), we passed the activation of the final layers y to $\log(1 + \exp(y)) + 1$ in order to make the beta distribution unimodal ($\alpha_\theta > 1$ and $\beta_\theta > 1$). 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 is used in the environment. The dimensions of output d are eight in the TRP environment and nine in the thermal regulation environment. The latter includes a one-dimensional evaporative action in addition to the eight-dimensional motor control of the quadruped robot.

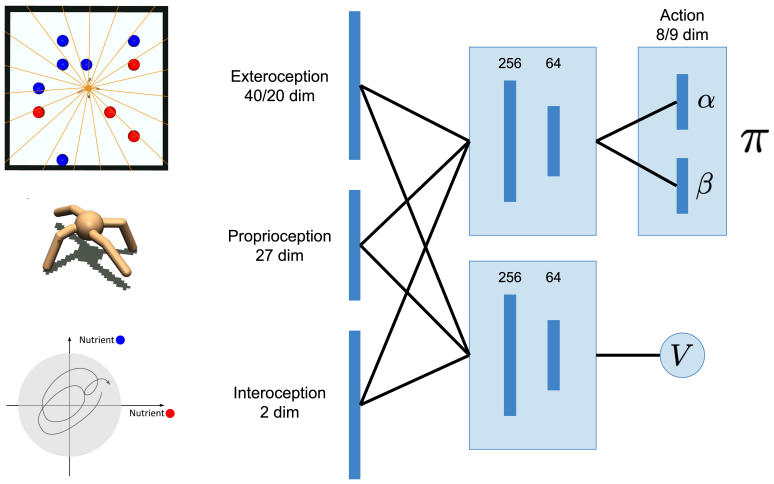

Figure S1: The architecture of the neural network used in our experiments. The agent's observations are composed of an exteroception x^e (the number of sensor dimensions is 40 for the TRP and 20 for the temperature control experiment), a 27-dimensional proprioception x^p and a two-dimensional interoception x^i (temperature and energy in the temperature control experiment).

The objective function to be maximized consists of four components:

$$J(\theta, \phi) = \hat{\mathbf{E}}_{\pi_{\mathrm{old}}} \left[L^{CLIP}(\theta) - c_1 L^{VF}(\phi) + c_2 S(\pi_\theta) - c_3 \tilde{D}(\pi_{\mathrm{old}} || \pi_\theta) \right], \tag{7}$$

where θ is the policy parameter, and ϕ is the value prediction parameter. We used the same hyperparameters, $c_1 = 0.5$, $c_2 = 0.001$, and $c_3 = 0.001$, throughout the experiments. The detailed definitions of each component are as follows.

$$L^{CLIP} = \min \left(\frac{\pi_\theta(u|x)}{\pi_{\mathrm{old}}(u|x)} \hat{A}_{\pi_{\mathrm{old}}}, \mathrm{clip} \left(\frac{\pi_\theta(u|x)}{\pi_{\mathrm{old}}(u|x)}, 1 - \epsilon, 1 + \epsilon \right) \hat{A}_{\pi_{\mathrm{old}}} \right), \tag{8}$$

$$L^{VF} = (V_{\mathrm{target}} - V_\phi(x))^2, \tag{9}$$

$$S(\pi_\theta) = H(\pi_\theta(\cdot|x)), \tag{10}$$

$$\tilde{D}(\pi_{\mathrm{old}} || \pi_\theta) = \log \frac{\pi_{\mathrm{old}}(u|x)}{\pi_\theta(u|x)}, \tag{11}$$

where $L^{CLIP}(\theta)$ is the surrogate loss of PPO for the policy improvement and $L^{VF}(\phi)$ is the value-prediction loss. $S(\pi_\theta)$ is the entropy bonus to enhance the exploration, where $H(\pi_\theta)$ is the entropy of the action given an observation. $\tilde{D}(\pi_{\rm old}||\pi_\theta)$ is the approximated Kullback-Leibler divergence penalty between the current and the previous policy $\pi_{\rm old}$, which is known to stabilise the optimisation (Hsu et al., 2020). The expectation operators in the above components were approximated by the stochastic gradient descent using minibatches.

We note that in optimisations, except for in the cart-pole setting, we did not provide the terminal ('done') signals to the agent and then reset the environment. This is because the terminal state value cannot be trivially defined in homeostatic reward definitions. The exception is the cart-pole setting, in which the terminal state values can be exactly introduced as zero.

We used the Adam optimizer with epsilon parameter 10^{-5} for optimisation with a learning rate that started from 3×10^{-4} and gradually decreased to 10^{-5} along with 500 training iterations. 3×10^5 training batch data were collected using ten worker threads in parallel, and a mini-batch size of 5×10^4 was used for the stochastic gradient descent. A summary of the hyperparameters is provided in Table S1. The algorithm was implemented on our customized version of the PFRL, a Pytorch-based deep RL package (Fujita et al., 2021).

Table S1: PPO hyperparameters used for the optimization

Hyperparameter	Value
Iterations	500
Batch size	3×10^5
Minibatch size	5×10^4
SGD update epochs	30
learning rate (initial)	3×10^{-4}
learning rate (final)	10^{-5}
Adam epsilon	10^{-5}
Maximum gradient norm	0.5
discount factor (γ)	0.99
GAE lambda (λ)	0.95
Clipping parameter (ϵ)	0.3
Value loss clipping parameter	10
Value loss coefficient (c_1)	0.5
Entropy coefficient (c_2)	10^{-3}
KL coefficient (c_3)	10^{-3}
Number of sampler threads	10

B Parameters of Reward Scaling and Bias

Table S2: Reward Scaling and Bias Parameters

Conditions	β_1	b
Homeostatic-shaped ($r_{\rm homeo} = \beta_1(D(x^i) - D(x^{i'}))$)	100	-
Homeostatic ($r_{\rm homeo} = -\beta_1 D(x^i)$)	0.1	-
Homeostatic-biased ($r_{\rm homeo} = -\beta_1 D(x^i) + b$)	1	0.1
Cart-Pole ($r_{\rm homeo} = -\beta_1$ if terminal, otherwise 0)	100	-

C Additional Details of the Thermal Regulation Environment

Our thermodynamic model of the agent is based on a model of the body temperature from a lizard (Porter et al., 1973; Fei et al., 2012). To incorporate the heat generated from the motor, we adopted a simple quadratic relationship with the motor output and the heat generation inspired by research

on the electric motor (Venkataraman et al., 2005). Our model of the core body temperature τ is described as:

$$C\frac{d\tau}{dt} = \delta Q(\tau, u, u_{\mathrm{ev}}), \tag{12}$$

where τ is the body core temperature of the animal and $C = C_a M$ is the heat capacity of the body. δQ is the amount of heat that is added to the body of the agent. 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. An identical step size with the decision step ($\delta t = 0.05$) was used in the simulation. δQ is composed of six components that are affected by environmental conditions and motor controls.

$$\delta Q(\tau, u, u_{\mathrm{ev}}) = \delta Q_{\mathrm{solar}} + \delta Q_{\mathrm{conv}} + \delta Q_{\mathrm{longwave}} + \delta Q_{\mathrm{cond}} + \delta Q_{\mathrm{m}} - \delta Q_{\mathrm{ev}} \tag{13}$$

Individual components are described as follows:

$$\delta Q_{\mathrm{solar}} = \alpha_L A_p Q_{\mathrm{solar}}, \tag{14}$$

$$\delta Q_{\mathrm{conv}} = h_L A_{\mathrm{air}} \left(\tau_{\mathrm{air}} - \tau\right), \tag{15}$$

$$\delta Q_{\mathrm{longwave}} = \epsilon_{\mathrm{skin}} A_{\mathrm{down}} \sigma_{SB} (\tau_{\mathrm{earth}}^4 - \tau^4) + \epsilon_{\mathrm{skin}} A_{\mathrm{up}} \sigma_{SB} (\tau_{\mathrm{air}}^4 - \tau^4), \tag{16}$$

$$\delta Q_{\mathrm{cond}} = \frac{A_{\mathrm{contact}} K_l \left(\tau_{\mathrm{earth}} - \tau\right)}{\Delta/2}, \tag{17}$$

$$\delta Q_{\mathrm{m}} = k u^\top u, \tag{18}$$

$$\delta Q_{\mathrm{ev}} = 0.5(u_{\mathrm{ev}}^{max} - u_{\mathrm{ev}}^{min}) \left(u_{\mathrm{ev}} + 1\right) + u_{\mathrm{ev}}^{min}. \tag{19}$$

Table S3 shows the parameters of the model. Values of the parameters are almost entirely adopted from the research that created the original model (Fei et al., 2012).

Table S3: Parameters of the Thermodynamics Model

Parameter	Value
Agent mass (M)	0.19
Agent heat capacity (C_a)	3762
Solar radiation (Q_{solar})	300
Skin absorbance (α_L)	0.936
Agent thickness (Δ)	0.015
Thermal conductivity (K_l)	0.502
Convection coefficient (h_L)	10.45
Agent shape coefficient (a)	0.0314
Agent area (A_L)	$\pi a M^{2/3}$
Projected agent area (A_p)	$0.4A_L$
Contacting area with the earth (A_{down})	$0.3A_L$
Areas of skin facing upward (A_{up})	$0.6A_L$
Skin area that is exposed in the air (A_{air})	$0.9A_L$
Area agent contacts with the ground (A_{contact})	$0.1A_L$
Emissivity of agent's skin (ϵ_{skin})	0.95
Stefan-Boltzmann constant (σ_{SB})	5.67×10^{-8}
Land temperature (τ_{earth})	303
Air temperature (τ_{air})	298
Maximum heat dissipation action (u_{ev}^{max})	0.3
Minimum heat dissipation action (u_{ev}^{min})	$0.272M$
Motor-heat coefficient (k)	5

D HOMEOSTATIC REINFORCEMENT LEARNING AS AN UPPER BOUND MINIMIZATION OF THE DIVERGENCE MINIMIZATION

We introduce the notation of the T-step sequence of observations $\bar{x} \triangleq x_1 x_2 \ldots x_T$. We assume the target distribution to be $P^*(\bar{x}) \triangleq \prod_{t=1}^{T} P^*(x_t)$ and the actual distribution realized by an agent with

strategy π to be $P_\pi(\bar{x})$. The KL divergence between P^* and P_π is transformed as

$$\frac{1}{T} D_{\mathrm{KL}}(P_\pi || P^*) = \frac{1}{T} \sum_{\bar{x}} P_\pi(\bar{x}) \log \frac{P_\pi(\bar{x})}{P^*(\bar{x})} \tag{20}$$

$$= \frac{1}{T} \sum_{\bar{x}} P_\pi(\bar{x}) \left[\log P_\pi(\bar{x}) - \log P^*(\bar{x})\right] \tag{21}$$

$$= -J_{\mathrm{h}} - \frac{1}{T} S(\pi) \tag{22}$$

$$\leq -J_{\mathrm{h}} \tag{23}$$

where $J_{\mathrm{h}} = \frac{1}{T} \sum_{\bar{x}} P_\pi(\bar{x}) \left[\log P^*(\bar{x})\right]$ and $S(\pi)$ is the entropy of the trajectory $-\sum_{\bar{x}} P_\pi(\bar{x}) \log P_\pi(\bar{x})$. And

$$J_{\mathrm{h}} = \frac{1}{T} \sum_{\bar{x}} P_\pi(\bar{x}) \left[\log P^*(\bar{x})\right] \tag{24}$$

$$= \sum_{\bar{x}} P_\pi(\bar{x}) \left[\frac{1}{T} \sum_{t=0}^{T} \log P^*(x_t)\right]. \tag{25}$$

This equality suggests that the maximization of the homeostatic objective through RL corresponds to the minimization of the KL divergence $D_{\mathrm{KL}}(P_\pi || P^*)$ from the upper bound. The distribution matching problem is more rigorously treated in recent imitation learning frameworks (Ghasemipour et al., 2020; Ke et al., 2020), and further technical advances are discussed in this field.

D.1 Temporal Difference of Drive Function as A Homeostatic Reward enhanced by Potential-based Reward-Shaping.

We will show that the homeostatic reward $r_{\mathrm{homeo}} = D(x_t) - D(x_{t+1})$ in the Markov decision process (MDP) can be derived from the reward $r = \log P^*(x)$ with a multivariate Gaussian assumption of $P^*(x)$, using reward transformations that preserve the optimal policy.

For simplicity, we introduce an equality $\overset{\pi}{=}$, which includes the scaling $r \overset{\pi}{=} \alpha r$, the baseline shift with a constant $r \overset{\pi}{=} r + b$, and the potential-based reward shaping $r(x, u, x') \overset{\pi}{=} r(x, u, x') + \gamma\Phi(x') - \Phi(x)$, where $\Phi(x)$ is the state-dependent potential function introduced by Ng et al. (1999).

Then, the reward based on the target distribution is transformed as

$$r_t = \log P^*(x_t) \tag{26}$$

$$\overset{\pi}{=} \log P^*(x_t) + \gamma\Phi(x_{t+1}) - \Phi(x_t) \tag{27}$$

We now introduce a potential function $\Phi(x) = \sum_{k=0}^{\infty} \gamma^k \log P^*(x) = \log P^*(x)/(1-\gamma)$. Subsequently, the reward is transformed as

$$r_t \overset{\pi}{=} \log P^*(x_t) + \gamma \frac{\log P^*(x_{t+1})}{1-\gamma} - \frac{\log P^*(x_t)}{1-\gamma} \tag{28}$$

$$\overset{\pi}{=} \log P^*(x_{t+1}) - \log P^*(x_t) \tag{29}$$

Finally, we assume that $\log P^*(x) \propto -D(x)$. Therefore, we can directly derive

$$r_t \overset{\pi}{=} D(x_t) - D(x_{t+1}) \tag{30}$$

$$= r_{\mathrm{homeo}} \tag{31}$$

In our main text, we assumed $D(x) = ||x^i - x^i_*||^2$. This corresponds to assuming that the target distribution is a multivariate Gaussian with a diagonal covariance ($P^*(x) = \mathcal{N}(x^i | x^i_*, \Sigma)$, $\Sigma = \sigma^2 I$ where $\sigma^2 > 0$ and I is the identity matrix).

Potential-based reward shaping can be regarded as the initialisation of the action value function using the potential Φ (Wiewiora, 2003). Our assumption $\Phi(x) = \log P^*(x)/(1-\gamma)$ may be a crude assumption because it assumes that the agent stays in the same state for an indefinite period. However, this shaping may provide a reasonable initialisation of the value function if the reward function is known and is smooth, as in our homeostatic RL settings.

E THE HOMEOSTATIC BEHAVIOR DOES NOT EMERGE FROM SIMPLE FOOD COLLECTION REWARDS

The agents were trained in the TRP environment with the same settings as in the other experiments. In this condition, the agent receives a reward of $+1$ for taking either red or blue food, and 0 otherwise. In this experiment, the agent received terminal information in the same manner as in the cart-pole setting. The figure shows the average (thick line) and individual results of the five runs (thin lines) with different random seeds.

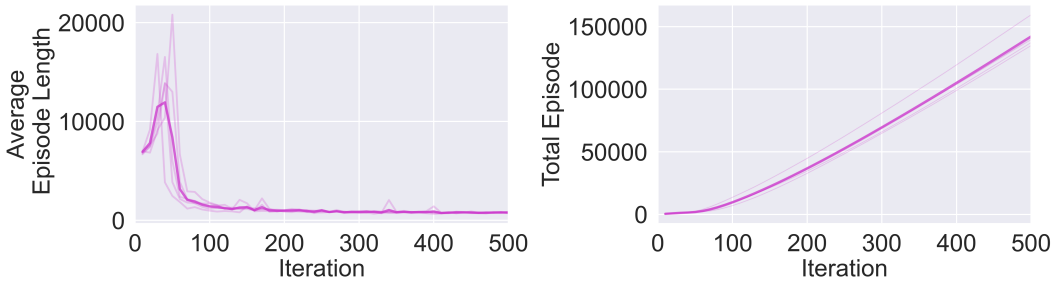

Figure S2: Performances of the optimization process with the food-collection reward condition.

F THE DETAILS OF THE BEHAVIOURAL PREFERENCE EXPERIMENT AND FURTHER DISCUSSION

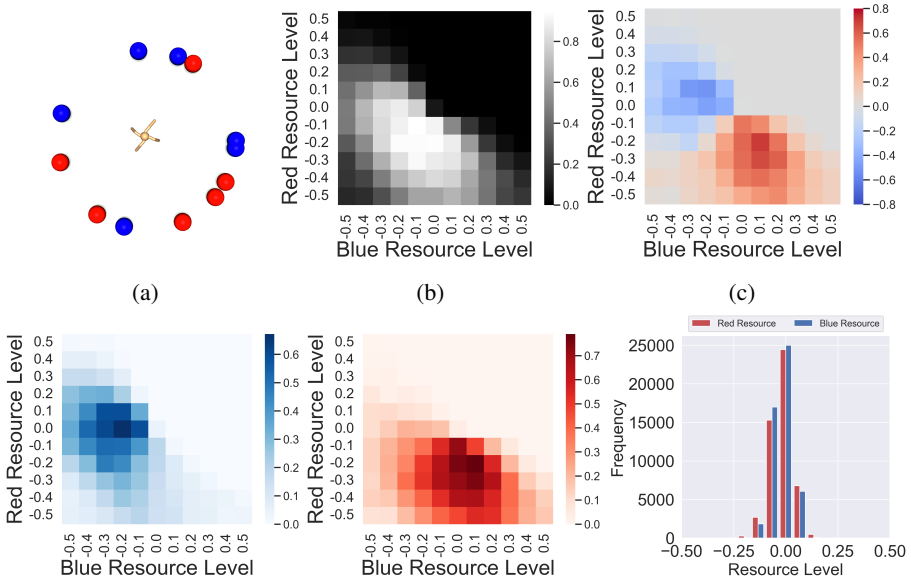

Figure S3: Behavioural experiment of the agent trained with (blue: 6, red: 4)-TRP+homeostatic-shaped setting. a) The overview of the initial condition of the experiment. The positions of red and blue ball clusters are randomly flipped between trials. b) Food collection tendency of the agent, depending on the internal state of the agent. c) Preference of the food collection in the experiment. d-e) Tendencies of the blue and red resource capture. f) Histogram of the internal nutrient states sampled in the TRP for 50,000 steps. Panel b to e are averaged results over five individually trained agents.

In this experiment, we manually clamped the agent's nutritional state at specific levels and observed the agent's choice of food resources afterwards. We found that agents were still active in this condition, and they changed their behaviours depending on their interoception. This experiment did not include the training process, and all the agent parameters were fixed.

The settings are shown in Figure S3a[2]. 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. We randomly located food resources to remove the locational bias in each trial because we observed that the agent used in this experiment tended to move toward a specific direction in the preliminary experiment. A single trial terminates if the agent consumes any one of the resources, or when 300 decision steps are passed. During the trial, the interoception was clamped to the value of interest. One hundred trials, $N = 100$, were conducted for each condition. Blue and red food consumption instances, N_{blue} and N_{red}, were counted and divided by the total number of trials N.

Figure S3 represents the average of the results from the five trained agents. Panel (b) shows the distribution where the agent captured any one of the red or blue resources $(N_{\mathrm{blue}} + N_{\mathrm{red}})/N$. This panel shows that the agent's preference of food resources was dependent on the interoception, which is obtained by $(N_{\mathrm{red}} - N_{\mathrm{blue}})/N$. Panel (d) and (e) are the food capturing rates of individual food resources N_{blue}/N and N_{red}/N.

As demonstrated in panels (b) and (c), we found that the tendency of the food-capturing behaviour decreased if the distance from the setpoint was large. We suspect that this performance degradation is due to the fact that the agent had a very biased experience after the learning process has progressed. Panel (f) is the histogram of the internal state during 50,000 steps of the agent sampled in the original TRP environment. We can confirm that both nutrient states are approximately in the range of $[-0.25, 0.25]$; agents might only be optimised around this distribution and thus they cannot generalise this knowledge beyond the experiences demonstrated here. Agents should be able to adapt their survival strategies appropriately in situations where they have little or no experience, which may be improved by using model-based learning with a world model (Ha & Schmidhuber, 2018; Hafner et al., 2020a) that includes the agent's physiological states.

G ADDITIONAL BEHAVIOURAL DATA

Figure S4 shows the behavioural preference for food capturing. The procedure is the same as that described in section F in the main text. In this experiment, the agent's internal state was clamped at values from -1 to 1 for each nutrient. We can observe that the agent stops the foraging behaviour if the nutrient deficits become large.

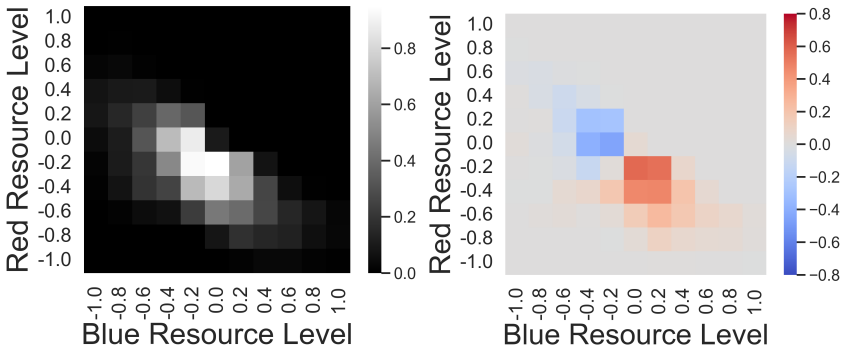

Figure S4: Food capturing preference dependent on the internal nutrient state. The result was averaged over five individually trained agents. (left) Food capturing rate. (right) Food preference depending on the agent's internal state.

[2]Panels (a-c) are identical to Figure 4 in the main text.

G.1 Behaviour Preferences of the Agent in (blue: 5, red:5)-TRP Environment

Figure S5 shows the results of the behaviour preference experiment with an agent trained in a (blue: 5, red: 5)-TRP environment. The optimisation process and the process for behavioural preference are the same as in the (blue: 6, red: 4)-TRP environment in Section F.

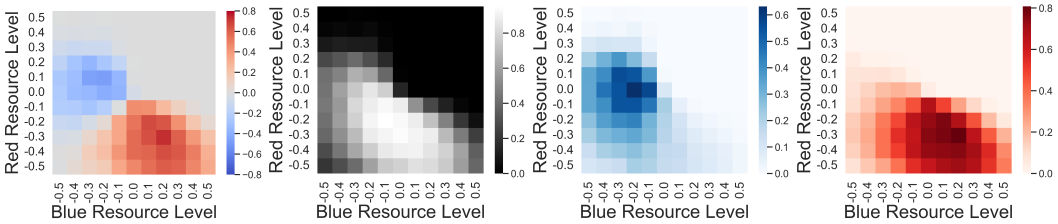

Figure S5: Food capturing preference of the agent trained in (blue: 5, red:5)-TRP. The result was averaged over five individually trained agents.

H Perturbation Experiment of the Thermal Regulation Environment

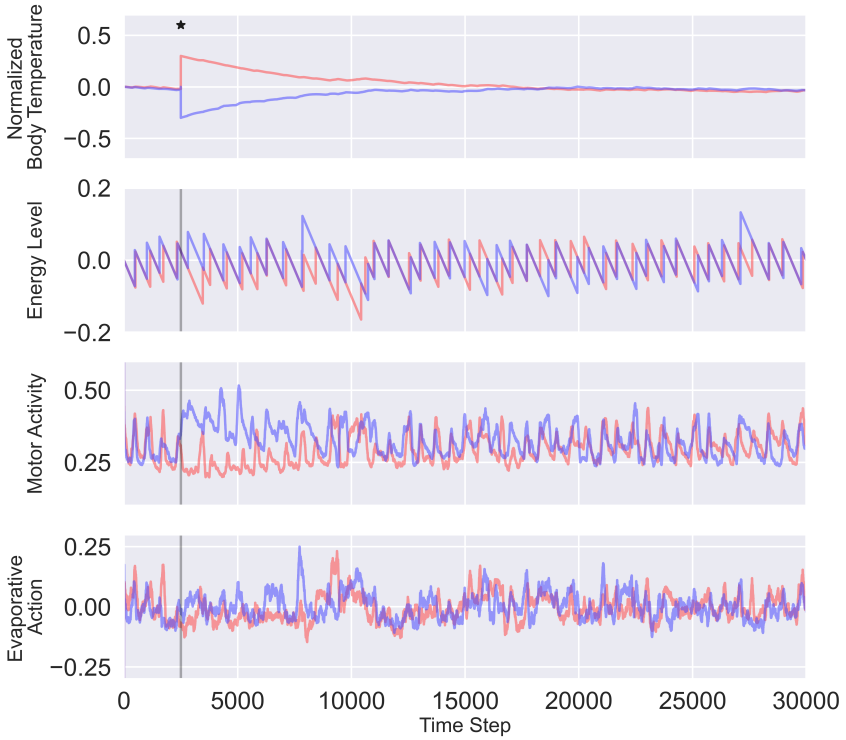

Figure S6: Results of the temperature-perturbation experiment. The agent (trained under the homeostatic-shaped condition) in environments (same seeds) receives the body temperature perturbation (± 0.3) at time step 2,500 (represented by a star and grey lines). The red line represents the positive perturbation and the blue line represents the negative counterpart (motor activities and evaporative actions are averaged using the last 100 steps for clarity). We can observe that the agent successfully returns to the setpoint ($\tau = 0$). In addition, following the body temperature change, we can observe that the agent regulates the food intake and motor activities.

I DETAILS OF THE STATISTICAL TEST OF BODY TEMPERATURE DEPENDENT MOTOR ACTIVITIES

We collected ten independent trajectories of motor activities for each of the three experimental settings. Each independent trajectory includes 5,000 steps. We clamped the thermal observation as 0.2 for the 'over-heat' setting and -0.2 for the 'freezing' setting. We clamped the thermal observation in the 'normal" setting. We used the parameter of the agent optimised through 500 PPO iterations with a homeostatic-shaped condition.

A total of 5,000 steps were averaged for each experimental setting. We confirmed the normality of the samples using the Shapiro-Wilk test. Because all of the experiment settings ('normal', 'over-heated', 'freezing') were not statistically different from the normal distribution ($p > 0.2$), we assumed the normality of samples in each category. We also tested for equal variance between 'normal'/'overheated' and 'normal'/'freezing' using the F-test. Because the statistical significance between 'normal' and 'freezing' was only slightly larger than 5% ($p \approx 0.06$), we assumed the variances were not equal and used Welch's t-test to evaluate the statistical significance between 'normal'/'overheated' and 'normal'/'freezing' conditions. We also evaluated effect sizes (Cohen's d) of the 'normal'/'over-heated' and 'normal'/'freezing' conditions, and these were found to be larger than nine.