Available online at www.sciencedirect.com

ScienceDirect

Physics of Life Reviews 31 (2019) 214–232

PHYSICS of LIFE reviews

www.elsevier.com/locate/plrev

Review

Neurocomputational theories of homeostatic control

Oliver J. Hulme [a], Tobias Morville [a], Boris Gutkin [b,c,*]

[a] *Danish Research Centre for Magnetic Resonance, Centre for Functional and Diagnostic Imaging and Research, Copenhagen University Hospital Hvidovre, Kettegard Allé 30, 2650, Hvidovre, Denmark*

[b] *Group for Neural Theory, LNC INSERM U960, DEC École Normale Supérieure PSL University, Paris, France*

[c] *Center for Cognition and Decision Making, Institute for Cognitive Neuroscience, NRU Higher School of Economics, Moscow, Russia*

Received 10 September 2018; received in revised form 5 June 2019; accepted 6 July 2019
Available online 19 July 2019
Communicated by Felix Schoeller

Abstract

Homeostasis is a problem for all living agents. It entails predictively regulating internal states within the bounds compatible with survival in order to maximise fitness. This can be achieved physiologically, through complex hierarchies of autonomic regulation, but it must also be achieved via behavioural control, both reactive and proactive. Here we briefly review some of the major theories of homeostatic control and their historical cognates, addressing how they tackle the optimisation of both physiological and behavioural homeostasis. We start with optimal control approaches, setting up key concepts, exploring their strengths and limitations. We then concentrate on contemporary neurocomputational approaches to homeostatic control. We primarily focus on a branch of reinforcement learning known as homeostatic reinforcement learning (HRL). A central premise of HRL is that reward optimisation is directly coupled to homeostatic control. A central construct in this framework is the drive function which maps from homeostatic state to motivational drive, where reductions in drive are operationally defined as reward values. We explain HRL's main advantages, empirical applications, and conceptual insights. Notably, we show how simple constraints on the drive function can yield a normative account of predictive control, as well as account for phenomena such as satiety, risk aversion, and interactions between competing homeostatic needs. We illustrate how HRL agents can learn to avoid hazardous states without any need to experience them, and how HRL can be applied in clinical domains. Finally, we outline several challenges to HRL, and how survival constraints and active inference models could circumvent these problems.

Keywords: Homeostasis; Alostasis; Homeostatic reinforcement learning; Active inference; Computational neuroscience

1. Homeostasis as a problem

Physiological states are insulated from hazardous fluctuations in the external environment by the operational and computational processes known as homeostasis [1–3]. These homeostatic processes unfold over deep hierarchies of

* Corresponding author.
E-mail address: boris.gutkin@ens.fr (B. Gutkin).

https://doi.org/10.1016/j.plrev.2019.07.005
1571-0645/

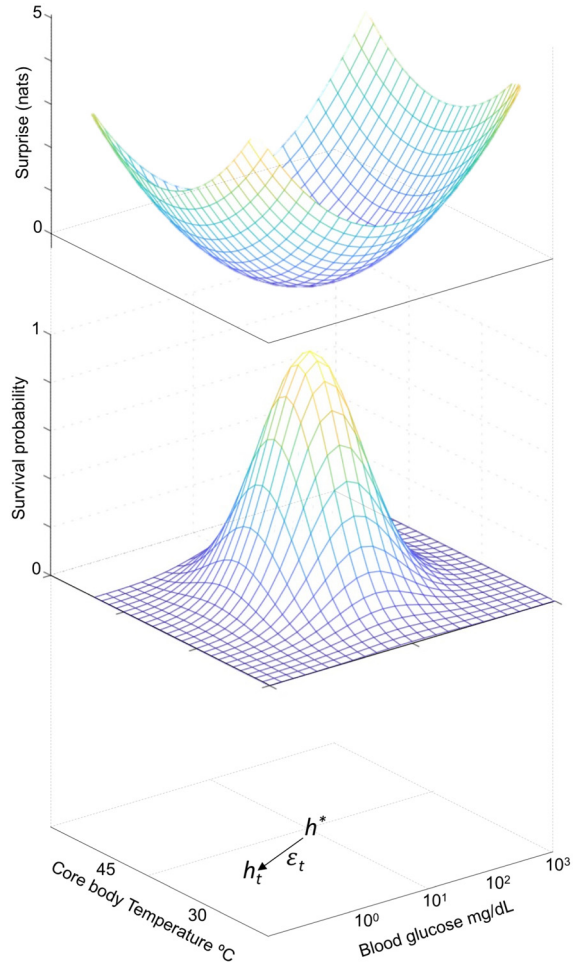

Fig. 1. Homeostatic state space, survival probability, and surprise. The lower tier shows a 2-dimensional homeostatic state space, where h^* denotes a homeostatic set point, h_t the current homeostatic state at time t, and the homeostatic error, defined as the Euclidean distance between them ε_t. The middle tier illustrates a homeostatic survival probability surface, depicted over the same state space, thus highlighting the relation between occupation of that homeostatic state and the conditional probability of survival (over some arbitrary time interval) in that state. This probability surface is an illustration, since actual survival probabilities have, to our knowledge, never been systematically inferred with any precision for higher organisms. Upper tier shows a homeostatic surprise surface, where surprise is defined as the negative log probability of observing the homeostatic state in an agent that is alive.

biophysical structure, molecules to networks of agents, and over a spectrum of timescales, milliseconds to years. Such physiological systems need to be highly regulated because the space of homeostatic states that support survival is so restricted. In other words, staying alive entails keeping to a small subset of highly frequented internal states, because excursions from those are punished either by death or reproductive failure. Indeed, in this regard, death can be defined circularly as the extreme and irreversible loss of homeostasis.

The stability of homeostatic states is thus prerequisite to the continuation of life. Stability here is coarsely analogous to the stability of mechanical equillibria, in which with the system returns to its equillibria point from small perturbations. These equillibria points are known in physiology as set points, which can themselves be perturbed, modulated, and controlled. Following from its definition above, one can think of death as a physiological perturbation that cannot be returned from. To take some common human examples, death reliably results from surprisingly narrow excursions in internal states from set point (Fig. 1a & b); for core body temperature it is an excursion from set point of $+5$ or -11 degrees Celsius; for blood osmolality it is $+/-25$ mOm/kg; and for blood glucose it is -2 mM or $+10$ mM outside of the normal range of 4–8 mM. Whilst these limits depend on the duration of the excursion and the joint trajectory of other homeostatic variables, it is a somewhat humbling fact that, surrounded by a vast ocean of pos-

sible mortality, all biological agents live on a tiny island of habitable internal states (Fig. 1b). From this perspective, the fitness afforded by any behavioural policy is determined (by definition) by its ability to steer the agent as close as possible to the trajectories of internal states that maximize survival.

The stability of physiological systems extends even to rheostasis (also known as allostasis), the dynamic process by which homeostatic set points themselves shift – for instance through stress, illness, or hibernation, through to longer scale circadian or circannual rhythms, developmental or reproductive phases [4]. Under these processes set points are still stable, with return to set points still occurring after perturbations, however the set points themselves maybe non-stationary over longer timescales, or under sudden changes of behavioural or autonomic priorities. This underscores the point that homeostatic stability does not necessarily entail homeostatic stationarity. The ultimate goal of homeostatic regulation is not to preserve the constancy of the internal milieu per se, but to continually adjust the milieu in order to maximize fitness [5]. In other words set points should always be defended, but they should also be dynamically adjusted. We will use the term homeostasis in its broadest sense to include this rheostatic component.

In any treatment of homeostasis, a useful but coarse distinction is between automated physiological processes – physiological homeostasis, and homeostasis mediated via overt behaviour – behavioural homeostasis. To contrast the two, consider what happens to an inactive organism as time passes. Basal metabolic needs will manifest in state variables either continually drifting, or becoming volatile. Physiological homeostasis cannot mitigate these excursions indefinitely, eventually its coordinated mechanisms are insufficient on their own to maintain physiological stability as deprivation sets in. The consequential increase in homeostatic error, defined as the distance of current state from the set point (Fig. 1, lower) can only reliably be rectified by some behavioural exchange with the environment. The behavioural control of homeostasis is thus a perpetual problem throughout the lifecourse for all motile agents. Such homeostatic control, if it is to be evolutionarily adaptive, consists of tracking, estimating, predicting and prioritizing trajectories of homeostatic errors as a function of the fitness that those trajectories afford. At bottom, this is a computational problem, and a deeply challenging one.

Our aim in this paper is to provide an overview of neurocomputational theories of homeostatic control addressing how they tackle this conjoint optimisation. We start with optimal control approaches, setting up key concepts, and expanding on their limitations. We then move onto more contemporary approaches, in particularly focusing on a recent branch of reinforcement learning known as homeostatic reinforcement learning (HRL). We explain its main advantages, empirical applications, and conceptual insights. We then outline some challenges to HRL and reinforcement learning in general, and how active inference models attempt to circumvent these problems.

2. A brief history of homeostatic theory

Early models of homeostasis were adapted from control theory, a mathematical branch of engineering that relied on exploiting varieties of feedback architectures [6]. In order to maintain steady-state equilibria, a controller (the brain) converts an input into a motor command, which is relayed to the plant (the body) resulting in a motor response, resulting in a new input. This idea of error correction encapsulates the logic of those early models, where correction was deployed to keep vital variables close to their set points. In a homeostatic context (Fig. 2a), the input to the controller would be a homeostatic error, which in turn generates a motor command, resulting in a behavioural exchange with the environment. This exchange may result in a modulation of the current physiological state, thus determining an update to the homeostatic error for the next timestep, that then inputs again onto the controller for iterative error correction. Notice that this is a reactive form of homeostatic control – the organism needs to necessarily experience the homeostatic error in order for the controller to react. A simple rendering of this in terms of energy homeostasis, would be where the homeostatic error is a glucosensory error, computed as the difference between current states (inferred from central and peripheral glucose sensors) and a euglycemic reference state (set point), where a glucosensory controller translates these errors into commands sent to the viscero-motor plant (for autonomic gluco-regulatory responses) and to the skeletal motor system (for overt behavioural responses such as foraging) until homeostatic error is minimized.

Generally, apart from being over-simplified for biological applications, direct feedback systems are observed to be noisy and unstable [6]. For example, delay or noise in the time taken from information to pass through the control system can lead to hunting, a phenomenon that results in the system perpetually oscillating. There are several ways to solve this problem. In systems biology, integral feedback control [7] relies on approximately the same architecture as direct feedback, but feedback is in the form of the time-integral of the output. As a result, the system only performs its regulatory action when its steady-state undergoes some perturbation. Typically, this is aimed at generating perfect and

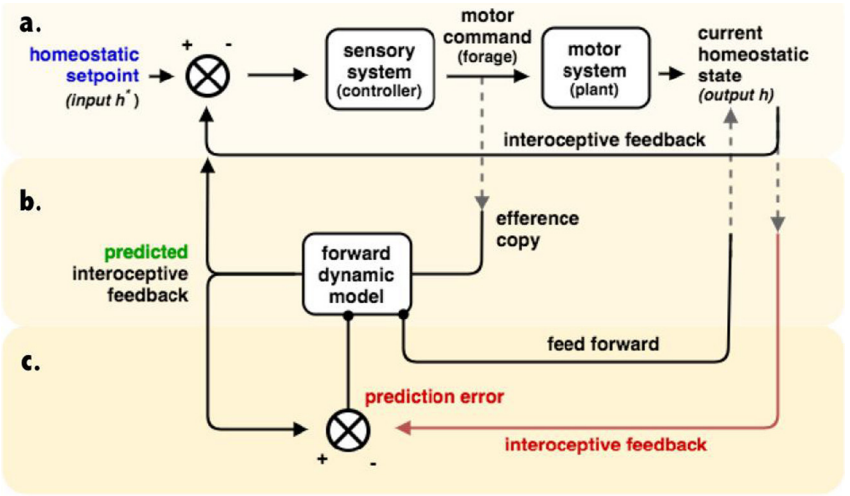

Fig. 2. Optimal control theory. **a**, the upper subsystem of this schematic depicts direct feedback control. This subsystem combines a controller (the sensory system), with a plant (the motor system) that performs actions in order to control the current homeostatic state. Homeostatic state is sensed by interoceptive input, which via comparison to set point results in a homeostatic error to be forwarded to the controller. Further commands iteratively work at minimising this homeostatic error. **b**, in the middle subsystem, a copy of the motor command is sent to the forward dynamic model which acts to predict the future interoceptive input, conditioned on the motor command having been executed. Any residual errors between the set point and this predicted state are again iteratively minimised with further commands. **c**, in the lower subsystem a prediction error, computed as the error between the current and the predicted state is used to update this forward dynamic model. Figure adapted from [6] also appears in pre-print [15].

robust adaptation, whereby the system variable returns to its baseline, after a sustained state change, irrespective of its magnitude. Whilst this might have desirable properties for a sensory system minimizing redundancy, it could have undesirable properties for homeostatic control, because survival probabilities are most likely stationary with respect to homeostatic state and thus the control structures need to retain this information by not adapting, in order to implement an effective survival optimisation.

Another control solution is to insert a forward dynamic module into the regulatory loop. Like direct feedback control, the controller receives a homeostatic error, from which it outputs a command for the plant, which acts on the variable of interest (Fig. 2b) to reduce the error. Unlike in simple direct feedback control, an efference copy is sent in parallel to the forward model, which predicts the viscerosensory state that will follow from the action, which is compared to the desired state. This results in a homeostatic prediction error that inputs back onto the controller, and this error is then recursively minimized over time by the same system. Importantly, introducing a forward model moves the control system into a regime of predictive processing, and thus these models can offer opportunities for explaining anticipatory control phenomena that cannot be explained by the purely reactive schemes (Fig. 2a). To keep the forward model calibrated, a prediction error between the current state and what was predicted by the forward model, is then used to update the forward dynamic model (Fig. 2c).

The forward models can work well if they reliably predicts future homeostatic errors, allowing the controller to act to minimize homeostatic error. However, these frameworks are challenged when confronted with the question of what actually constitutes the homeostatically rational behaviour that the commands should prescribe. These models assume that the controller already knows what commands to execute under which conditions. However in most ecological settings, this is not possible, because there are few (if any) stationary mappings between homeostatic errors and homeostatically rational behaviours. This is especially the case if the external and internal environments are uncertain. An even deeper problem with these types of model is that homeostatic error minimization per se, is not necessarily the overarching objective of homeostatic control, since homeostatic errors are scaled arbitrarily by setting the units of measurement. This becomes apparent when focusing on the problem of how to prioritize the urgency of actions that remediate different bundles of homeostatic error, say how to arbitrate between minimizing one unit of hydration error, and three units of energetic error, versus three hydration units and one energetic unit. As suggested in the lower tier of Fig. 1, one might simply compute a homeostatic error as the Euclidean distance from set point, and optimise action to minimize that error. However, it can be shown that, under this strategy, changing the units of measurement

inherently imposes an arbitrary re-prioritization of homeostatic dimensions. The more challenging question is then how to motivate prioritization between homeostatic dimensions in a way that is biologically plausible as a strategy for fitness maximisation.

Drive reduction theory (DR), proposed by Clark Leonard Hull, was arguably the first theory to directly apply ideas of homeostasis to motivation and behaviour [8,9]. On DR, instead of directly minimising the homeostatic error, primary drive was proposed as an overarching minimandum, minimized over the long-run, guiding biological agents to select actions that promote survival. The probability of a given action (the reaction potential), is then determined by the product of habit strength and drive. Drive-reducing actions are reinforced into habits, which provides a means by which behaviour minimizes homeostatic error. Drive here is a negative valenced state that the agent works to attenuate, and in so doing, this attenuates the associated homeostatic deficits that cause it. In this framework, a drive function is a mapping between the homeostatic space and motivational drive. Inherent in the evolutionary logic of DR, though only implicitly formulated, is a calibration of drive to the statistics of survival, such that homeostatic states are afforded drive as a function of their survival hazard: "*...when any of the commodities or conditions necessary for individual or species survival are lacking, or when they deviate materially from the optimum, a state of primary need is said to exist*." [9], noting that need here is synonymous with drive.

By crude analogy to mechanical systems, drive can be thought of as the potential energy that a simple mechanical system minimises. If evolution works to select phenotypic drive functions to match the homeostatic realities of survival, then it should set drive minima to approximate homeostatic equillibria points (set points) where survival probabilities are highest, and it should ensure stable as opposed to unstable equillibria points, which like mechanical systems it can engineer via its second derivative. Concave drive functions would be unstable, whereas convex will be stable. Thus, the form of the drive function may have deep connections to the stability of internal states.

Though conceptually insightful as formal models of simple control, direct feedback models have shortcomings that made their candidacy for explaining agent-level adaptive behaviour short-lived. Those shortcomings were also to plague DR theory, which was criticised for its inability to explain certain empirical observations: first, animals develop drives before any homeostatic errors have developed [10] such as eating when still satiated [11], drinking when still hydrated [12], and even shivering before becoming cold [13,14]; second, its mechanistic account of learning was generally acknowledged as being poorly predictive of behaviour, even in narrow experimentally controlled conditions.

3. Homeostatic reinforcement learning

Despite the move away from DR theory, some of these problems on how to learn behavioural patterns under general environmental feedback were addressed in the following decades outside of this literature. Notably, a formal framework for learning called reinforcement learning (RL), addresses some of those shortcomings by effectively combining behavioural psychology and optimal control theory. There is an extensive literature on the various RL approaches in modern artificial intelligence, machine learning economics, and engineering [16]. Here we will focus on those algorithms that are most biologically relevant. As a basic tenant for RL algorithms to work, a state space for the agent must be defined and should have the Markov property. These states can be explicitly observable by the agent or partially probabilistically signalled to the agent, or inferred. In a large class of RL algorithms (so-called model free) the structure of the state and action space is not known by the agents; in other words, the agent does not compute a structural model of the state or action space. Second, RL posits that actions and states are endowed with "values" or utilities that translate disparate outcomes of actions into a common currency. This allows them to be compared when generating a policy, and it is these values that are learned and retained by an RL agent. Finally, learning is driven in RL toward reward maximization, typically it is the sum of expected future discounted rewards that is maximized. In general the learning of values is driven by a signal that is a function of the received reward and the expected reward. Most relevant for biology is the family of algorithms based on the Rescorla-Wagner model, where the teaching signal is obtained from the difference between the expected reward value and the received reward value. This has been extended by temporal difference learning algorithm to incorporate updates to reward predictions prior to rewards being actually experienced [17]. The reward prediction error signals of this algorithm show a remarkable consilience with the phasic signals of the dopaminergic system [18].

It should thus be noted in this context that RL generally asserts the existence of rewards and how they are distributed in the environment, rather than ground reward in terms of drive, homeostasis, or survival. This approach

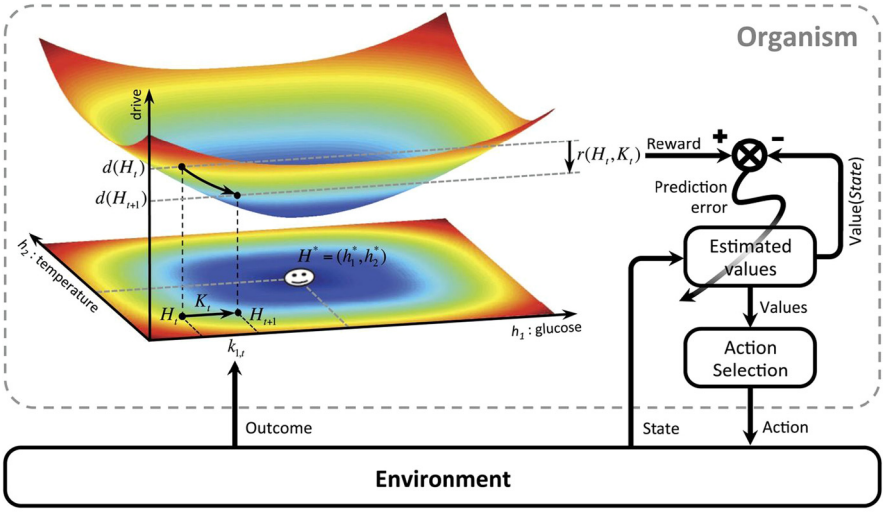

Fig. 3. Homeostatic reinforcement learning. Upper left, the surface shows a putative drive function, mapping from homeostatic state space to drive. In comparing the estimated reward value to the actual reward experienced (with negative reward defined as drive inflation), a reward prediction error is computed, which then updates future value estimates. Behavioral Actions are selected as a function of these estimated values. Note that $d(H_t)$ in this figure is identical to $D(H_t)$ in this paper. Adapted from [22] with permission.

ultimately entails designing algorithms aimed at maximizing cumulative reward over specific temporal horizons. In any environment endowed with some temporal regularity, maximizing expected future reward, necessitates RL algorithms capable of anticipatory action. The RL policies to solve this anticipatory problem thus have an important bearing on the problems of anticipatory responding for DR.

From this perspective it makes sense to bridge between reinforcement learning and DR theory. DR theory offers a foundational definition of reward, which RL lacks, and RL provides a sophisticated suite of tools by which actions can be learnt to maximize reward, which DR lacks. Indeed HRL, achieves this by inserting a drive-reduction definition of reward into a toolbox of standard RL algorithms [19,20]. The non-linear shape of the drive function determines how drive changes for any action, given the current homeostatic state. This replaces the arbitrary concept of reward in RL, by defining drive with respect to homeostatic state, and by defining reward as the drive reduction that attends homeostatic error reduction. The HRL framework shows (in eqs. (5)–(9)) that an agent minimizing drive will maximize reward, which minimizes homeostatic error. Hence, HRL directly couples homeostatic optimisation to reward maximisation. Interestingly HRL is able to replicate some of the anticipatory responding phenomena that had originally troubled the original DR theory [19] (see eqs. (11)–(17)). This can be considered a special case of optimal control theory, which shows that for any environment with temporal structure, acting predictively on expected future deviations instead of reactively, can be more effective in minimising long-run error [21]. Thus, a simple interpretation of this would be, that an agent whose superordinate objective was to minimise drive in the ways described by Hull, would be more effective in that minimisation by deploying anticipatory strategies. Ironically, HRL shows that the anticipatory control issues that where historically problematic for DR theory, are in fact mandated for effective drive-reduction.

3.1. Homeostatic state space and the drive function

The point of departure for HRL is to define a homeostatic state space as a multidimensional metric space where each dimension represents one regulated physiological variable (the horizontal plane in Figs. 1b & 3). For simplicity the different homeostatic dimensions are assumed to be independent of each other and to be Markovian. The homeostatic state of the agent can thus be represented via its position in this N-dimensional space, denoted by $H_t = (h_{1,t}, h_{2,t}, .., h_{N,t})$, where $h_{j,t}$ represents the state of the j-th homeostatic variable at time t. For example, $h_{j,t}$ can refer to the animal's glucose level, or body temperature, and so forth. The homeostatic set point, defined as the survival optimal internal state (assuming that this is stationary, and thus suppressing subscript t), is denoted by $H^* = (h_1^*, h_2^*, .., h_N^*)$. In calling it a set point we make no wider assertion that it should be stationary over time, though for parsimony of exposition, in many of our examples we will assume it to be. As a mapping from homeostasis state to

motivational state, we construct a drive function. We want this function to have the following properties: to be convex, and hence to have a single minima to align with homeostatic optima; to be well behaved – at least twice differentiable, continuous almost everywhere on the domain; to be finite on finite sub-domains; and to have the properties of a distance metric. A natural choice is a generalization of the Minkowski distance measures with the following functional form (depicted as the vertical dimension in Fig. 3):

$$D(H_t) = \sqrt[m]{\sum_{j=1}^{N} \left| h_j^* - h_{j,t} \right|^n}, \tag{1}$$

where, $D: \mathbb{R}^n \rightarrow \mathbb{R}$. Parameters m and n are free, and can be chosen to according to considerations of the biological plausibility of behaviours this engenders (discuss this further in the section below, 4.3 *Constraining the drive function*).

3.2. Defining reward value

Having defined drive, we can now provide a formal definition for reward, that is conceptually derived from DR. Assume that as the result of an action a_t the animal receives an outcome o_t at time t. The impact of this outcome on different dimensions of the animal's internal state can be denoted by $K_t = (k_{1,t}, k_{2,t}, .., k_{N,t})$. For example, $k_{j,t}$ can be the quantity of glucose consumed. Such an outcome will result in a transition of the homeostatic state from H_t to $H_{t+1} = H_t + K_t$ and consequently, a transition of the drive state (Fig. 3) from $D(H_t)$ to $D(H_{t+1}) = D(H_t + K_t)$. Accordingly, the reward value of this outcome can be defined as the consequent reduction in drive:

$$\begin{aligned} r(H_t, K_t) &= D(H_t) - D(H_{t+1}) \\ &= D(H_t) - D(H_t + K_t). \end{aligned} \tag{2}$$

Intuitively, the reward value of an outcome depends on the ability of its components in reducing the homeostatic distance from the set point. This homeostatically defined reward value can then be incorporated into any RL algorithm to estimate the values of the states and actions to be taken. As a paradigmatic example of an RL algorithm we start with Q-learning. Here the key learning term is the reward prediction error signal, δ_t (Fig. 3, right). This signal is computed each time the agent takes an action and experiences an outcome from its environment. This prediction error is calculated by comparing the prior expected value $Q(s_t, a_t)$ of taking action a_t starting from a state s_t, and the realized value after receiving reward r_t:

$$\delta_t = r_t + \gamma \cdot \max_{a_{t+1}} \left\{ Q(s_{t+1}, a_{t+1}) \right\} - Q(s_t, a_t), \tag{3}$$

where maximum Q-value of all feasible actions a_{t+1} available at state s_{t+1} is discounted by the temporal discounting factor $\gamma \in (0, 1)$. Note that state s_t is more general than homeostatic state in the sense that it can represent environmental and agent-level states, where the latter can include both cognitive and homeostatic states. This prediction error signal is hypothesized to be reported by the phasic firing of midbrain dopamine neurons [18]. This signal can be used to update the estimated value of actions

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \cdot \delta_t, \tag{4}$$

where α is the learning rate, representing the degree to which the prediction error iteratively adjusts the Q-values. We should point out that homeostatically defined reward can in principle plug into any RL algorithm.

3.3. Reward maximisation via drive minimisation

The HRL theory as formulated above is normative, in the sense that the policies that maximize the sum of future discounted reward, are the same policies that minimize future discounted homeostatic errors. In other words, reward maximization and homeostatic control are two sides of the same coin. In order to demonstrate this, we note that the behavioural policies result in a sequence of outcomes that alter the internal state of the agent, and thus can be evaluated as a function of their trajectory through homeostatic state space. We define $\mathcal{P}(H_0)$ as the set of all homeostatic trajectories that start at H_0, and end at H^*. Hence the homeostatic consequences of a policy can be described as a

homeostatic trajectory, denoted by $p = \{K_0, K_1, \dots\}$, an ordered sequence of N-dimensional homeostatic states, resulting in a total of w instances of homeostatic states and ending at the optimal point H^*. Each K_i is an N-dimensional vector, determining the length and direction of a homeostatic transition caused by each outcome. Here we assume that $p_i \in \mathcal{P}(H_0)$ is a sample homeostatic trajectory consisting of $w-1$ transitions between homeostatic states.

Let us now make two key definitions: the sum of discounted drives (SDD) and the sum of discounted rewards (SDR).

For each homeostatic trajectory p_i we can define *SDD* as the sum of discounted drives through that trajectory:

$$SDD_{p_i}(H_0) = \sum_{t=0}^{w-1} \gamma^t \cdot D(H_{t+1}). \tag{5}$$

Similarly, we define *SDR* as the sum of discounted rewards through trajectory p_i:

$$\begin{aligned} SDR_{p_i}(H_0) &= \sum_{t=0}^{w-1} \gamma^t \cdot r_t \\ &= \sum_{t=0}^{w-1} \gamma^t \cdot \big(D(H_t) - D(H_{t+1})\big) \end{aligned} \tag{6}$$

As a trajectory p_i denotes the ordered sequence of homeostatic states: $\{H_0 = H_{i,0}, H_{i,1}, H_{i,2}, \dots, H_{i,w} = H^*\}$. For simplicity, denoting $D(H_{i,t})$ by $D_{i,t}$, drive for trajectory p_i will evolve in the following sequence: $\{D_0 = D_{i,0}, D_{i,1}, D_{i,2}, \dots, D_{i,w} = D(H^*) = 0\}$. Thus, we can rewrite eq. (5) to express the sum of discounted drives as:

$$SDD_{p_i}(H_0) = D_{i,1} + \gamma \cdot D_{i,2} + \gamma^2 \cdot D_{i,3} + \cdots + \gamma^{w-1} \cdot D^*. \tag{7}$$

This can be expressed as the sum of discounted rewards:

$$\begin{aligned} SDR_{p_i}(H_0) &= r_{i,0} + \gamma \cdot r_{i,1} + \gamma^2 \cdot r_{i,2} + \cdots + \gamma^{w-1} \cdot r_{i,w-1} \\ &= (D_0 - D_{i,1}) + \gamma \cdot (D_{i,1} - D_{i,2}) + \gamma^2 \cdot (D_{i,2} - D_{i,3}) + \gamma^{w-1}\big(D_{i,w-1} - D^*\big) \\ &= D_0 + (\gamma - 1) \cdot SDD_{p_i}(H_0). \end{aligned} \tag{8}$$

Since D_0 is fixed for all policies, and $(\gamma - 1)$ is negative, it can thus be concluded that if a certain trajectory from $\mathcal{P}(H_0)$ maximizes $SDR(H_0)$, then it will also minimize $SDD(H_0)$, and vice versa:

$$argmin_{p \in \mathcal{P}(H_0)} SDD_p(H_0) = argmax_{p \in \mathcal{P}(H_0)} SDR_p(H_0). \tag{9}$$

Notice that this is independent of the form of the drive function as long as the trajectories that minimize it exists and can be found.

3.4. Temporal discounting

We should note that the temporal discounting factor plays a key role in this derivation. Had this factor been set to unity (meaning no discounting of rewards in the future) then the value of any given behavioural policy would depend only on its start and end point in homeostatic state space. Hence, any policies that had the same endpoints but embodied different homeostatic trajectories would have the same value, and would thus be path independent. This is biologically untenable, since trajectories which incur large excursions to hazardous homeostatic states are necessarily dangerous for the organism's survival, and thus any adaptive motivational system should be able to score these trajectories as an approximate function of their fitness prospects. As a corollary, with temporal discounting, HRL allows the agent to learn behavioural policies that keep it away from hazardous homeostatic states. In other words, temporal discounting of rewards over time may be an adaptation that was mandated to ensure homeostatic stability, and thus survival.

4. What HRL has achieved

As formulated above, HRL has a number of properties that align well with observed behavioural phenomena, as well as providing an unifying explanatory framework. We begin with behavioural phenomena before proceeding to applications in the domain of addiction.

4.1. State-dependent valuation

One interesting phenomena observed behaviourally is that animals are able to seemingly adjust their valuation of outcomes according to their current or anticipated homeostatic states, even if the same outcome has never been experienced in such states. For example, in rats that have previously experienced a salty-flavoured solution, induction of sodium appetite enhances their preference for the salt-associated flavour, even though they have never experienced a sodium deficiency before [23,24]. This shift in preference can be generalized, such that induction of a hunger state in an animal can energise both their food seeking and water seeking habitual responses. HRL suggests that this outcome independent energizing effect is an approximate way of updating the value of state-action pairs when the homeostatic, and thus motivational, state shifts instantaneously. Assuming that the animal is trained under the fixed homeostatic state H, and then tested in a novel internal state H', the Q-values can be approximated in the new motivational state by

$$Q_1(s,a) = \frac{D(H')}{D(H)} \cdot Q_0(s,a), \tag{10}$$

where $Q_0(s,a)$ represents action-values learned by the habitual system after the training period, and Q_1 represents the updated Q-value. According to this update rule, all the prepotent actions will be energized if deviation from the homeostatic set point increases in the new homeostatic state, whether or not the outcome of those actions are more desired in the new state. This value adjustment equation is optimal only when the updated state is a scalar multiple of the old state, such that $H' = c \cdot H$. Thus, having this value update allows the agent to estimate the reward associated with an outcome in a state-dependent manner without having to have previously experienced the outcome in this specific state before. This in turn automatically biases actions away from those likely to incur hazardous excursions (see [25] for further discussion). Interestingly this state dependent modulation of value can be seen as a model for incentive sensitization [26] having multiplicative form [27].

4.2. Anticipatory control

HRL as defined above yields an important property for the policies that are learned in order to defend homeostasis. Notably, HRL leads to optimal policies that can be either pre-emptive, corrective, or some combination of the two (as opposed to only being corrective). Pre-emptive action is paramount for the slow dynamics of most physiological systems, as well as for exploiting complex environments, in which corrective actions alone are not efficient in delivering long-run homeostasis. Given that rewards are temporally discounted, HRL will generate policies that take a pre-emptive action upon cues that predict future homeostatic excursions (see [25] for a simulated example). A demonstration of this anticipatory faculty is as follows. Let us consider a single internal homeostatically variable regulated over time h_t with a set point value h^*. And let us consider that when an animal is at h_t, a sensory cue predicts a future homeostatic challenge l that perturbs h_t further away from h^*. For simplicity, and without loss of generality, assume that this challenge occurs two time-steps into the future. Similarly, for clarity of presentation, we will cap the future horizon to a total of three time-steps. Let us assume that the animal has learned to take a pre-emptive action a that results in an outcome moving the internal state by a distance of k. For simplicity, we suppress time notation in a, k, and l. We can now explore what the optimal action should be. Here we define optimality as a function of minimizing *SDD*. Note again, that with temporal discounting in place, under the HRL framework, minimizing *SDD* entails maximizing *SDR*. Here we explore what outcome will maximize the *SDR*.

Starting from h^*, taking the pre-emptive action a following the predictive cue, changes the homeostatic state by k such that homeostatic state evolves as follows:

$$h_0 = h^* \rightarrow h_1 = h^* + k. \tag{11}$$

The reward value of this change k is then given by

$$r_1\big(h^*, h^* + k\big) = D\big(h^*\big) - D\big(h^* + k\big). \tag{12}$$

The homeostatic challenge l will cause the homeostatic state to change as

$$h_1 = h^* + k \rightarrow h_2 = h^* + k - l. \tag{13}$$

The reward value of this challenge is thus

$$r_2\big(h^* + k, h^* + k - l\big) = D\big(h^* + k\big) - D\big(h^* + k - l\big). \tag{14}$$

Finally, at the final timestep homeostatic state returns to h^*, so the final reward value is

$$r_3\big(h^* + k - l, h^*\big) = D\big(h^* + k - l\big) - D\big(h^*\big). \tag{15}$$

From this setup, we can now compute the *SDR* for action a for the three timesteps:

$$\begin{aligned} SDR_a &= r_1\big(h^*, h^* + k\big) + \gamma \cdot r_2\big(h^* + k, h^* + k - l\big) + \gamma^2 \cdot r\big(h^* + k - l, h^*\big) \\ &= -\sqrt[m]{|k|^n} + \gamma\big(\sqrt[m]{|k|^n} - \sqrt[m]{|k - l|^n}\big) + \gamma^2\big(\sqrt[m]{|k - l|^n}\big). \end{aligned} \tag{16}$$

We can find the extrema of the SDR by examining the zeros of the derivative of the SDR with respect to l, since *SDR* is convex in l with a single maximum. By construction, *SDR* will attain its maximum when its derivative with respect to l_x is zero. Taking the derivative of SDR_a with respect to l, we show that we get the following equation:

$$\begin{aligned} &(\gamma - 1)k^{\frac{n}{m}-1} = \gamma(\gamma - 1)\big(l - k^{\frac{n}{m}-1}\big) \\ &k = l\frac{1}{1 + \gamma^{\frac{m}{m-n}}}. \end{aligned} \tag{17}$$

Since γ is always upper bounded by 1, we see that k should always be some fraction of l. As temporal discounting slows (as γ tends to 1) this fraction tends to 1/2. As discounting becomes more rapid (as γ tends to 0) this fraction tends to 1. In other words, the optimal action for countering expected homeostatic challenges should be to move homeostatic state by between 50% and 100% of the size of the impending challenge, depending on the level of temporal discounting. How temporal discounting is optimised will be the focus of future work.

4.3. Constraining the drive function

Let us now impose a biologically plausible constraint on the drive function: we will assume that the exponents adhere to the following inequality: $n > m > 1$ [20]. A drive function with this constraint yields a geometry that is convex everywhere, which elicits a number of interesting properties for behavioural homeostasis, and that may have important ramifications for decision-making and reward. We detail four such properties in the following subsections.

4.4. Excursion aversion mandates risk aversion

In economics, the curvature of an agent's utility function – the function that maps from outcomes to subjective value – is predictive of, and indeed equivalent, to the agent's risk preference. In terms of the resulting choices, in the domain of gains, a concave utility functions yields risk-aversion, whereas a convex utility function yields risk-seeking. In this case, since the reward value function is twice differentiable almost everywhere (since this is true for the drive function), concavity is achieved if:

$$\frac{d^2 r(H_t, K_t)}{dk_{j,t}^2} < 0: \quad \text{for } k_{j,t} > 0. \tag{18}$$

Under the HRL definition of reward, the constrained drive function satisfies this condition:

$$\frac{\partial^2 r(H_t, K_t)}{\partial k_{j,t}^2} = \frac{\partial^2 (D(H_t) - D(H_t, H_t + K_t))}{\partial k_{j,t}^2} = (-1)\big|h_t^* - h_{j,t} - k_{j,t}\big|^{\frac{n}{m}-2}\frac{m}{n}\left(\frac{m}{n} - 1\right). \tag{19}$$

As long as $n > m > 1$, the second derivative above is negative, and thus risk aversion for gains will be expressed. Note that this is true for outcomes that increase and decrease homeostatic deviations, and thus risk aversion is predicted for both homeostatic gains and losses. Note that the same condition of concavity at the point of homeostatic equilibrium, is what determines it to be a stable equilibrium point. This points to a potentially deep connection between risk aversion and the stability of homeostasis that merits further empirical attention.

4.5. Excursion aversion mandates loss aversion

An interesting corollary is that, because of the constant curvature of the drive function, there is an inequality between the reward value of a unit increases in homeostatic error (a loss) and the reward value of unit decreases in homeostatic error (a gain). Where the utility of a loss exceeds that of the utility of the same sized gains is known in economics and psychology as loss aversion [28]. Letting K_t^+ be a homeostatic error decreasing outcome, and K_t^- it's additive inverse, acting to induce an equally sized increase in homeostatic error. Loss aversion is demonstrated by the fact that:

$$\left|r\left(H_t, K_t^+\right)\right| < \left|r\left(H_t, K_t^-\right)\right| : \quad \text{for } K_t^+ < \left(H_t - H^*\right). \tag{20}$$

In other words, the loss of K carries more disutility (negative reward value) than the gain of K carries utility (positive reward value). This provides a normative basis for why loss aversion should be observed within behavioural homeostatic settings. To our knowledge this has not been formally tested.

4.6. The excitatory effect of homeostatic deprivation

Again, under this constrained drive function, we can show that the reward value of a unit reduction of homeostatic error grows with increasing deviations from the set point. This is a form of alliesthesia, the changing of reward value (or pleasure) as a function of homeostatic state [29]. To show this, we take the first derivative of the reward function with respect to deviations from set point:

$$\frac{dr(H_t, K_t)}{d|h^* - h_{j,t}|} > 0 : \quad \text{for } k_{j,t} > 0. \tag{21}$$

The fact that this derivative is positive means that the reward value of a unit of homeostatic error reduction, increases with size of the initial homeostatic error. This formalises the idiom "*hunger is the best sauce*".

4.7. Inhibitory effect of irrelevant drive

It is interesting to note that under this constrained drive function, the reward value of an outcome is suppressed by deviations in "*irrelevant drives*", i.e. in directions orthogonal to the direction in the homeostatic space along which the action-outcome is taken:

$$\frac{\partial r(H_t, K_t)}{\partial |h^* - h_{b,t}|} < 0 : \quad \text{for all } b \neq j \forall k_{j,t}, h_{j,t} > 0. \tag{22}$$

This formalises the phenomena whereby one motivational dimension can inhibit other motivations [30].

4.8. Avoiding hazardous homeostatic excursions

In model-free RL, in order for an agent to learn the value of a state-action pair, the agent must experience this pair along with its associated outcomes. This is problematic for homeostatic agents that should not deviate far from set point, as such excursions are hazardous for survival. How does HRL resolve this problem? We have previously shown that an agent learning to control it's homeostatic state with a drive function constrained as above, will learn to avoid enacting policies which result in hazardous excursions without having to experience them [20]. For ease of exposition, lets us consider an agent on a one-dimensional state space which is naive to the value of any of its actions, and thus has an equal probability of moving toward or away from its set point. The probability of choosing to step away from the set point z times in a row decreases exponentially as z increases: $p(z) = 2^{-z}$, thus the larger the excursion the more improbable: for example, the probability of choosing stepping away once is $2^{-i} = 0.5$, the probability stepping away twice is $2^{-s} = 0.25$, and so on. Thus, it is highly likely for the agent to return at least one step back, before getting too far from its starting point. Recall that the step back will be associated with a positive reward. Further, when the agent returns to a state it had previously experienced, going in the same deviation-increasing direction will be less likely than in the first experience of the state, since the agent has already experienced the punishment caused by that state–action pair once. One can show that HRL agents in this setup result in homeostatic occupancy probabilities that

concentrate around the set point and with a rapidly vanishing probability of large deviations. This illustrates how HRL allows agents to explore local parts of their homeostatic state space, and to learn the value of the paths back toward set point, without having to experience hazardous excursions.

4.9. Reward prediction errors

It follows from the excitatatory effect of homeostatic deprivation (subsection 4.7), that reward values for homeostatic error reductions should decrease with decreasing levels of homeostatic error. If reward prediction errors are encoded by phasic signalling of midbrain dopaminergic cells, HRL predicts that these signals should attenuate with decreasing levels of homeostatic error (for the homeostatic dimension of the reward in question). In other words, as satiety for a homeostatic dimension increases, the reward prediction errors for this dimension should decrease in magnitude. There is some initial support for this prediction. In experiments where behavioural conditioning was driven by the animal's appetite for sodium, the phasic dopamine release into the nucleus accumbens, as measured by fast cyclic voltammetry, progressively decreased as the animal became satiated [31,32].

5. Applications of homeostatic reinforcement learning

As a fundamental framework for modelling how reward and decision-making depends on the internal states of the organism, it is hoped that HRL can have applications beyond basic biological science. So far, we have applied the framework to metabolic disorders [25] as well as the psychopathology of addiction [33]. Here we summarise the work done so far on addiction and related phenomena.

5.1. Alcohol tolerance

We previously showed that a learnt anticipatory response to the hypothermic effect of alcohol can explain classical experimental results on acquired alcohol tolerance [13,14]. In these experiments the animal is exposed to cold stimuli under acute alcohol injections whilst tracking core temperature. Alcohol induces a larger body temperature deviation in response to cold stimulus, than that seen under saline injections. With repeated exposure this alcohol-dependent effect diminishes, hence the experiment was interpreted as measuring progressive development of tolerance to the effect of alcohol. We showed that internal regulation of reward as predicted by HRL, together with a simple model for the acute tolerance response to alcohol, can account for the experimental observations, where the animal learns to increase its anticipatory response to mitigate the effect of alcohol (see [25]). Interestingly, the model also captured the transient increase in the body temperature in catch trials where the alcohol injection was omitted.

5.2. Cocaine addiction

The homeostatic perspective might also provide insight into addictions of substances that are not naturally subject to homeostatic regulation. Specifically HRL has been used to explain the escalation of cocaine seeking and consumption in rats with extended access to the drug [33]. In these experiments, animals were given access to cocaine either for a short period per daily session (1 hour per day) or a long access (6 hours per day). While the short access animals reached a stable daily dose and cadence of cocaine, the long access animals progressively escalated their daily doses and increased consumption frequency. We posited that escalation is due to an allostatic deviation of the set point that results from a progressive build-up of a regulated internal variable, induced by the persistent presence of the drug. We then developed an extended version of HRL, where an acute injection of cocaine produced a modelled pharmaco-dynamic dopamine response that we treated as a reward signal. We conjectured that the regulated variable is a function of the tonic dopamine levels, which in turn controls that location of the set point in the homeostatic state space and has a slow relaxation dynamic. While the exact nature of this regulated variable still remains to be determined, we further postulated that the regulated variable should specifically depend on the postsynaptic effects of the tonic dopamine outflow in the nucleus accumbens (and/or ventral striatum). Notably, our review of the literature lead us to propose that the tonic activation of D1 dopaminergic receptors on the D1-expressing NAcc neurons that project to the ventral pallidum, is a likely (but possibly not the sole) regulated homeostatic variable compatible with our theory (see [33] on the possible substrates).

In simulating instrumental cocaine-taking tasks, we showed that the HRL agent, when exposed to cocaine for a limited time per day, learns to consume the drug at a stable, and relatively limited dose. This happens because the slow tonic dopamine level relaxes to control levels over night and the set point remains stable over the long time scale. Under prolonged access, the HRL agent escalates the dose. This happens because the tonic level of dopamine does not have time to relax and builds up on this longer time scale. This induces the set point so as to create an increasing homeostatic error. The agent in turn learns to seek and consume progressively larger doses of the drug to pre-emptively counter these increasingly large deviations in the homeostatic state space. Hence HRL predicts that escalation is a learnt instrumental response, conditioned by the acute effect of cocaine, and the longer-term opponent process dependent on the tonic dopamine levels. Interestingly the allostatic mechanisms of this HRL model captured a number of observations associated with escalation experiments: the loading phase at the beginning of each session, the regular drug seeking subsequently, the increase in injection frequency as the per-injection dose was decreased [34,35]. Most importantly, the model could account for dose-induced relapse to drug seeking after a prolonged withdrawal. Such relapse could be seen only in a "model-based" version of HRL, where the agent explicitly tracked if it is in a drug-available (drug-influenced) or drug-free state. This lead us to speculate that certain central aspects addictive behaviours may be goal directed, as opposed to automatic and habitual [36].

6. Discussion

6.1. Challenges to HRL

The progress of HRL offers promise in many regards, though it is worth discussing some of the residual problems with these approaches, some of which are inherited from optimal control theory. Firstly, HRL is still untethered from any biological maximandum, relying instead on the drive functions which are chosen as a sensible approximation based on the behavioural and economic phenomena. As reasonable as this might be, it is fitting the drive function to the behaviour, which suffers the same circularity at the heart of economic conceptions of utility, and the behaviour-centric definitions of reward value. The circularity lies in the fact that drive function is fitted to behaviour, and then behavioural optimality is defined with respect to drive minimisation. In economics the utility functions are fitted to behaviour, and then behavioural optimality, or rationality, is defined with respect to utility maximisation. One plausible solution to this problem for homeostatic control problems, would be to specify the drive functions as a function of survival probability, which we will discuss in the next section. Secondly, almost all of the optimal control theory based schemes, which includes RL and thus HRL, are based on the Bellman equation (Hamilton-Jacobi-Bellman in continuous time), where the optimal policy that maximises the value function is the solution to the Bellman equation [37]. This assumes that the agent has access to hidden states with certainty, and that the optimal value function can be solved for all possible states [38]. Third, such learning schemes are notoriously slow, requiring hundreds if not thousands of trials to learn even the simplest value-function for a given setting [38]. Arguably, this could be mitigated by introducing a model-based learning scheme to the HRL. Here the organism would learn not only the homeostatically signalled value function, but also infer the structure of the external environment and internal state space transitions. In this case the drive function maybe dynamically modified as a function of the learned environmental or interoceptive cues, for instance signalling acute physiological dangers, the proximity of hazardous states, or of the availability of new reward modalities and actions. In [33] we took initial steps in exploring this direction, considering how the drive function (and thus the value function) would shift under environmental manipulation where drug outcomes were available or not. Yet much more work remains to be done in this direction. The limitations outlined here motivate the augmentation of HRL, whilst also considering the merits of other approaches. Here we discuss how these challenges can be addressed by other models, or by imposing additional constraints on HRL.

6.2. Active inference

Explaining the many and varied control processes that range from basic physiological homeostasis to active planning, and execution of pre-emptive actions arguably requires a global theory of neural structure and function. Active inference (AI) offers such a framework, resting on the assumption that the brain is a hierarchical prediction machine that seeks to maximise the evidence for its own model of the world, by minimising an upper bound on surprise. We will unpack this last sentence: "*maximising evidence for its own model*", entails simply maximising its probability

of existence, where the model is taken as the agents phenotype – in other words maximising its survival probability. "*...surprise*", on the other hand, is defined as the negative log probability of observing a given state (Fig. 1c). We think this has a natural interpretation in the context of homeostatic regulation. We opened this paper with the problem of staying alive as a problem of occupying the homeostatic states that afford high probability of survival. This can be restated in terms of minimising surprise, where the states that have a low probability of being observed, have a low probability precisely because they are the states that cause the agent to die, and thus carry high surprise (Fig. 1b & c). Minimising surprise, thus entails maximising survival, and thus fitness. So how is this surprise minimisation achieved according to AI?

According to AI, the brain embodies a hierarchy of multiple nested hypotheses about the world that it inhabits, and surprise is minimised at each level by minimising the discrepancy between incoming sensory signals and top-down predictions. In other words surprise is minimised by minimising prediction errors. Importantly, there are two ways prediction error, and thus surprise, can be minimised. The agent can engage in perception, in which it updates its predictions to conform to the sensory input. Or it can engage in action, in which it acts on the world, to change its sensory inputs to better match its sensory predictions. Both of these faculties are evidently relevant to homeostatic control. In the case of perception, for the agent to engage in homeostasis, it must engage in interoception in order to infer what homeostatic state it is in, and with what uncertainty. Note that this interoception is an inference problem in the same way that exteroception is. Both are subject to uncertainty, since in neither case can the hidden states of the environment, or the body be directly accessed. In the case of action, homeostatic control also becomes a highly relevant case. The agent is postulated to have genetically and epigenetically specified predictions for what homeostatic states it expects to occupy, from which prediction errors are computed by comparison to current interoceptive states. To reduce these prediction errors, and thus minimise surprise, the agent needs to act on the world, to engage in behavioural exchange in order to minimise these errors across its hierarchical networks.

In AI, a generative model establishes a probabilistic mapping between hidden causes (internal or external states) to their predicted sensory consequences (i.e. exteroceptive or interoceptive inputs). The generative model combines a prior with a likelihood function probabilistically mapping from hidden internal states to the interoceptive inputs. The generative model here is generative in the sense it can generates sensory predictions. Under AI models of homeostatic control [2,39], homeostatic set points are recast as prior belief distributions of the interoceptive states that the agents predicts it will occupy. These prior belief distributions are either genetically inherited, epigenetically or experientially programmed, and subject to allostatic modulations by other states over multiple timescales. Under evolutionary pressure, these genetic and epigenetic priors are selected to best fit the hypothesis of survival in the organisms expected ecological niche (see [38] for discussion). In other words, prior beliefs that do not fit well to the ecological niche, will result in agents being motivated to occupy states that afford low survival probability, and thus those priors (as phenotypes) will be selected against. Importantly, these prior beliefs replace the need for an explicit value function such as the drive function discussed above [40]. The value function such as the drive function in DR or HRL attracts the agent to the valuable states, of low drive, and in turn to states that obtain low homeostatic errors. The same is achieved in AI, but by virtue of acting to fulfil prior beliefs over homeostatic states. The trajectories of homeostatic state (p_i as defined in the HRL model) that support survival, would thus have a high value in an optimal control theoretical sense, are now defined as trajectories (of states) that are likely to occur, and thus have low surprise. Indeed taking the path integral of surprise scores the cumulative survival probability of that homeostatic trajectory, such that minimising it, maximises cumulative survival probability. If the agent is equipped with prior beliefs that are adaptive to its ecological niche then acting to sample viscero-sensory inputs that conforms to low surprise, and thus expecting to keep homeostasis error minimised becomes a self-fulfilling prophecy [39].

Stephan et al. [2] formulate a minimal model of hierarchical Bayesian homeostasis and its allostatic modulation. In Fig. 4 we reproduce this model to demonstrate how dynamics of key elements of AI (homeostatic state, surprise and action) evolve under environmental flux. In column a, environmental flux pushes homeostatic state away from its set point (encoded as a prior belief distribution over viscerosensory states) which results in a homeostatic prediction error, and the execution of a subsequent action in order to restore the homeostatic state to its set point. In column b, the agent expects a future deviation, and engages in allostatic control by modulating its prior beliefs distribution, in essence changing its beliefs as to what viscerosensory states its expects now to occupy. This results in the agent executing action to change its current visceral state in order to minimise deviation from set point, via interoceptive prediction error minimisation. Notice that because the deviation is expected, prediction error and subsequent action is smaller compared to column a. For this specific model, the interested reader is recommended to consult [2], or for

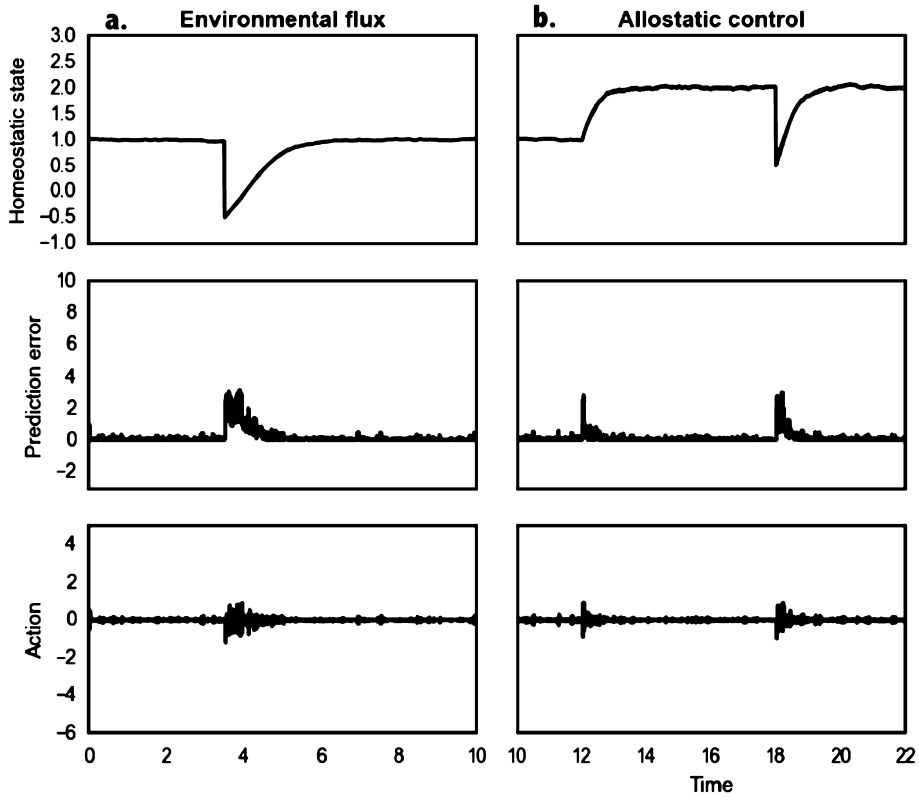

Fig. 4. Allostatic regulation and homeostatic control using hierarchal Bayes. See main text for details and [2] for modelling details. **Column a**, The effect of an environmentally evoked change in a single homeostatic variable (top), with resulting viscerosensory prediction errors generated (middle), and the resulting homeostatic action to remediate the errors (bottom). **Column b**, Depicts the same set of events but where the agent engages in allostatic control in which a future environmental flux is anticipated ahead of its occurrence. Set point moves ahead of the environmental flux (top), resulting in a lower (path integral of) prediction error (middle) and action (bottom) required.

a more general introduction to AI models see [41]. For a fuller treatment of the comparison between AI models and HRL see [15].

6.3. Survival constraints on drive

It should be noted at this stage, that there are some salient parallels between AI and DR theory. If we take seriously Hull's statements about drive minimisation being an optimisation of survival probability, then one can consider what drive function would be predicted in relation to survival probabilities. Defining drive in the same way surprise is defined, yields some interesting qualitative features:

$$D(H_t) = I(H_t) = -ln\big(Pr(H_t)\big), \tag{23}$$

where I is the surprisal of observing state H_t. Firstly, since survival probabilities are monotonically decreasing with respect to homeostatic error, drive is monotonically increasing with respect to increasing homeostatic errors, which is qualitatively what is specified by DR. Minimising drive thus minimises homeostatic error, where drive is highest for the most hazardous states, and lowest for the least hazardous. Secondly, if agents are to minimise drive over homeostatic trajectories, then they should minimise the integral of drive over time. A drive function specified as in eq. (23), means that minimising its time integral, would have the effect of maximising the cumulative survival probability over that integral. This can be seen in the following discrete example, where cumulative survival probability from the present up to time t is

$$S_t = \prod_{u=0}^{t} \phi_u = exp\left(\sum_{u=0}^{t} ln(\phi_u)\right), \tag{24}$$

which can be expressed as a sum over the log survival probabilities, where ϕ_u is the survival probability at time u. Similarly, we can define the cumulative drive CD, again from present up to time t, as a sum over (negative) surprise values, which is approximated by the sum over drives:

$$CD = exp\left(-\sum_{u=0}^{t} I(h_u)\right) = exp\left(\sum_{u=0}^{t} D_u\right). \tag{25}$$

As such, minimising cumulative drive is an approximate means of maximising cumulative survival:

$$\underset{p \in P(H_0)}{argmin}\ CD_p(H_0) \approx \underset{p \in P(H_0)}{argmax}\ S_p(H_0),$$

where $P(H_0)$ is the set of all homeostatic trajectories that start at H_0 and end anywhere in homeostatic state space H_t. Consideration of these drive trajectories is an important exercise, since agents should not just maximise survival probabilities over the narrowest of temporal windows, but over the temporal scales that contribute to their fitness prospects. For instance, for an agent to realise its fitness, it must at least reach sexual maturity. If such an agent was endowed with the drive function in eq. (23), then if it was successful in minimising the time integral of drive up to sexual maturity, then it would be maximising a major component of its fitness. Under these assumptions, the optimal drive function is fixed by the phenotype-specific mappings from homeostatic state to probability of survival. Intriguingly, if homeostatic survival probability functions are governed by central limit theorem, then the survival optimal drive functions are quadratic with respect to homeostatic error, and this mandates all of the key properties described for HRL above: risk aversion, loss aversion, the inhibitory effect of irrelevant drive, and the excitatory effect of homeostatic deviation. This is important since it provides a basis for grounding the drive function not on a fit to behaviour, but on a fundamental level constrained by the statistics of homeostatic survival. This offers an escape from the circularity of value. Furthermore, this perspective provides an intriguing insight that seemingly disparate behavioural phenomena might be emergent as evolutionary solutions to homeostatic control.

6.4. Drive without survival constraints

If the agent is to act to minimise any superordinate variable such as drive, then if the phenotypes that manifest as drive functions are subject to evolutionary selection pressure, then those phenotypes should be selected out as a function of their divergence from their surprise function. If an agent's drive function were to have any other functional form, then it would by definition be suboptimal with respect to survival. In other words, drive functions vary in fitness as a function of the degree to which they approximate the negative log survival probability, because this allows drive minimisation when evaluated over time, to optimise cumulative survival probability. To illustrate, we can entertain what would happen if this was not the case. An agent is endowed with a phenotype whose drive minima (the minima of its drive function with respect to homeostatic state) is such that it is motivated to visit say thermal states that are incompatible with survival (e.g. body temperature of 52 °C). The agent is destined to visit, or try to visit, fatal homeostatic states; the candidate phenotype thus has low fitness. For the same reason, an interoceptive system that is inefficient or biased in its representation of internal state, will result in motivations to visit suboptimal states and thus also have low fitness.

6.5. Survival statistics

This perspective provokes several questions, the most immediate being, how does the organism know the relationship between its homeostatic state and its chance of survival, i.e. how does the organism become a 'good' model of its environment (internal and external)? Learning about the relation between one's homeostatic state and the probability of death is difficult, not least because obtaining the first data point directly is game over. A gods-eye observer, observing species over evolutionary timescales, could in principle piece together such a survival function. But how does a single agent obtain this information if it cannot infer from exteroceptive input in its own lifetime? We postulate that

the relationship between homeostatic state and the probability of death is 'learnt' by an evolutionary process, where the optimal drive functions are encoded in the genotype (which scaffold epigenetic and developmental systems) and are subject to evolutionary optimisation. Recall this is in effect the same assumption embedded in AI theories of homeostatic control, where the prior belief distribution for interoceptive states is genetically, epigenetically, or developmentally programmed. This perspective of evolution offering optimal solutions to the "hypotheses of life" is congruent with contemporary theories of self-organisation, such as the free energy theory [38,42], escort evolutionary game theory [43] and other diffusion and flow-based accounts of evolution [44].

6.6. Empirical tractability

According to the perspective offered above, there should be a systematic relationship between survival statistics of homeostatic state, and the prior beliefs encoded and enacted by the brain. Insofar as it is possible to approximate the natural statistics of homeostatic survival, an experimenter can (in principle) approximate the set points one would normatively expect to observe under evolutionarily equilibria. However, for ethical reasons, high precision experimental data is not so feasible for larger lifeforms. For simpler lifeforms though, experimental data already exists; the best example to date is that afforded by the C. elegans lifespan machine, a modified document scanner used to accurately record time of death of a large population of C. elegans [45]. The possibility to record actuarial metrics such as force of mortality over time, with respect to homeostaticaly relevant variables such as temperature is now automated for large populations, yielding the means by which to estimate survival probability functions experimentally and with the precision of large ensembles. A more powerful and immediately applicable insight comes from realising that the specific functional form of prior beliefs fundamentally restricts the nature of expected interoceptive signals under homeostatic equilibria. As a result, the framework specifies qualitative normative predictions for the dependencies between homeostatic state, reward value, and behaviour.

6.7. The natural statistics of homeostasis

Restricting ourselves to the case study of glucose or energy homeostasis, we can ask what do these statistics look like? Serum glucose levels that deviate from the norm (less than 2.8 mmol/L) can cause cognitive impairment [46], seizure, coma [47], and ultimately death [48]. This is corroborated by data that shows correlation between all-cause mortality rates in humans and increased deviation from set point glucose, defined as the highest occupancy frequency state [49,50]. The supralinear relation between homeostatic error and mortality is not limited to serum glucose, and has been documented across species for other fundamental homeostatic variables, such as temperature [51] and osmolality [52]. This points to a limited set of attracting homeostatic states that are occupied with a high probability, and is commensurate with organisms entertaining low-entropy prior beliefs, as theories such as AI prescribe. These high frequency states are often organised around the mean of the distribution, which typically have the lowest force of mortality (the instantaneous rate of death). Taking this at face value means that the adaptive animal expects interoceptive signals that are consistent with a small set of set points and that increasing the distance of one's current (or expected future) states to those set points increases interoceptive surprise supralinearly.

6.8. Future

Though it is often lamented that in the biological sciences, there is much more data than theory, in the domain of homeostatic control, it is quite clear we have the converse problem. We have a set of theories that require discipline, falsification, and provocation from data. In this landscape it is a concern that theories of homeostatic control are not sufficiently falsifiable, a position for which we have some sympathy. Process theories are much needed to connect from the normative to the neural implementation. However, even at this nascent stage, there are ways in which the theories espoused here could be shown to be empirically wrong. Behavioural and neural data could, in principle at least, show that reinforcement and reward signalling is homeostatic-state invariant. Regularities in homeostatic control behaviour may prove to violate the phenomena we asserted were important features of a drive function, for instance if biological agents were not risk averse with respect to both homeostatic deviations and reductions. Being wrong or right is not the only way to judge the merit of a theory. We assert that being wrong in ways that provoke interesting experimental questions is a critical function of theoretical frameworks. Thus a good test of whether HRL and its cognates is a good

framework do not necessarily rest only on whether it is vindicated empirically, but also on whether it provokes fruitful avenues of experimentation, or generates new models and frameworks. As we articulated above, what is lacking is precise and systematic data on the homeostatic statistics of higher animals living in natural settings, on their survival statistics, and on their relation to choice behaviour and reward signalling, with concurrent homeostatic measurement. All of these domains of empirical data are currently sparse to non-existent. We thus end with an optimistic call to encourage ethologists, physiologists, psychologists and computational biologist to talk to each other and embark on charting this fundamental frontier.

Author contributions

All authors contributed equally to the writing. BG & OH worked primarily on the mathematical formalisms. TM worked primarily on the figures. All authors edited the work and approved the final version.

Declaration of Competing Interest

The authors declare no competing financial interests.

Acknowledgements

O.J.H (Lundbeck Foundation, ref: R140-2013-13057; Danish Research Council ref: 12-126925) T.M. (Lundbeck Foundation ref: R140-2013-13057). B.S.G acknowledges partial support from LABEX ANR-10-LABX-0087 IEC and from IDEX ANR-10-IDEX-0001-02 PSL*. This work was supported by the HSE Basic Research Program and the Russian Academic Excellence Project "5-100".

References

[1] Cannon WB. The wisdom of the body. New York, NY, US: W W Norton & Co; 1932.
[2] Stephan KE, Manjaly ZM, Mathys CD, Weber LAE, Paliwal S, Gard T, et al. Allostatic self-efficacy: a metacognitive theory of dyshomeostasis-induced fatigue and depression. Front Human Neurosci 2016;10. https://doi.org/10.3389/fnhum.2016.00550. 1–27.
[3] Sterling P, Laughlin S. Principles of neural design. MIT Press; 2015.
[4] Mrosovsky N. Rheostasis. USA: Oxford University Press; 1990.
[5] Sterling P. Physiology & behavior. Physiol Behav 2012;106:5–15. https://doi.org/10.1016/j.physbeh.2011.06.004.
[6] Carpenter R. Homeostasis: a plea for a unified approach. Adv Physiol Educ 2004;28:180–7. https://doi.org/10.1152/advan.00012.2004.
[7] Somvanshi PR, Patel AK, Bhartiya S, Venkatesh KV. Implementation of integral feedback control in biological systems. Wiley Interdiscip Rev, Syst Biol Med 2015;7:301–16. https://doi.org/10.1002/wsbm.1307.
[8] Hull CL. A behavior system; an introduction to behavior theory concerning the individual organism. New Haven, CT, US: Yale University Press; 1952.
[9] Hull CL. Principles of behavior: an introduction to behavior theory. Appleton-Century-Crofts; 1943.
[10] Bolles RC. Theory of motivation; 1975.
[11] Reppucci CJ, Petrovich GD. Learned food-cue stimulates persistent feeding in sated rats. Appetite 2012;59:437–47. https://doi.org/10.1016/j.appet.2012.06.007.
[12] Gawley DJ, Timberlake W, Lucas GA. Anticipatory drinking in rats: compensatory adjustments in the local rate of intake. Physiol Behav 1988;42:297–302. https://doi.org/10.1016/0031-9384(88)90086-8.
[13] Mansfield JG, Cunningham CL. Conditioning and extinction of tolerance to the hypothermic effect of ethanol in rats. J Comp Physiol Psychol 1980;94:962–9. https://doi.org/10.1037/h0077824.
[14] Mansfield JG, Benedict RS, Woods SC. Response specificity of behaviorally augmented tolerance to ethanol supports a learning interpretation. Psychopharmacology 1983;79:94–8. https://doi.org/10.1007/BF00427791.
[15] Morville T. The homeostatic logic of reward. bioRxiv 242974, https://doi.org/10.1101/242974, 2018.
[16] Sutton RS, Barto AG. Introduction to reinforcement learning. 1st ed. Cambridge, MA, USA: MIT Press; 1998.
[17] Sutton RS, Barto AG. Introduction to reinforcement learning. 1st ed. Cambridge, MA, USA: MIT Press; 1998.
[18] Schultz W. A neural substrate of prediction and reward. Science 1997;275:1593–9. https://doi.org/10.1126/science.275.5306.1593.
[19] Keramati M, Gutkin BS. A reinforcement learning theory for homeostatic regulation. In: Shawe-Taylor J, Zemel RS, Bartlett PL, Pereira F, Weinberger KQ, editors. Advances in neural information processing systems 24: 25th annual conference on neural information processing systems 2011. Neural Information Processing Systems (NIPS). ISBN 9781618395993, 2011. p. 82–90.
[20] Keramati M. Collecting reward to defend homeostasis: a homeostatic reinforcement learning theory. Biorxiv 2014:1–41. https://doi.org/10.1101/005140.
[21] Narendra KS, Annaswamy AM. Stable adaptive systems. Courier Corporation; 2012.

[22] Keramati M, Gutkin B. Homeostatic reinforcement learning for integrating reward collection and physiological stability. eLife 2014;3:e04811. https://doi.org/10.7554/eLife.04811.
[23] Berridge KC, Schulkin J. Palatability shift of a salt-associated incentive during sodium depletion. Q J Exp Psychol B 1989;41:121–38.
[24] Fudim OK. Sensory preconditioning of flavors with a formalin-produced sodium need. J Exp Psychol, Anim Behav Processes 1978;4:276–85.
[25] Keramati M, Gutkin B. Homeostatic reinforcement learning for integrating reward collection and physiological stability, https://doi.org/10.1101/005140, 2014.
[26] Robinson TE, Berridge KC. The neural basis of drug craving: an incentive-sensitization theory of addiction. Brains Res Rev 1993;18:247–91. https://doi.org/10.1016/0165-0173(93)90013-P.
[27] Zhang J, Berridge KC, Tindell AJ, Smith KS, Aldridge JW. A neural computational model of incentive salience. PLoS Comput Biol 2009;5:e1000437. https://doi.org/10.1371/journal.pcbi.1000437.g004.
[28] Kahneman D, Tversky A. Choices, values, and frames. Am Psychol 1984;39:341–50. https://doi.org/10.1037//0003-066X.39.4.341.
[29] Cabanac M. Physiological role of pleasure. Science 1971;173:1103–7.
[30] Dickinson A, Balleine B. The role of learning in the operation of motivational systems. In: Stevens' Handbook of Experimental Psychology, vol. 37; 2002. p. 407–19.
[31] Fortin SM, Roitman MF. Physiological state tunes mesolimbic signaling: lessons from sodium appetite and inspiration from Randall R. Sakai. Physiol Behav 2017;178:21–7. https://doi.org/10.1016/j.physbeh.2016.11.021.
[32] Fortin SM, Roitman MF. Challenges to body fluid homeostasis differentially recruit phasic dopamine signaling in a taste-selective manner. J Neurosci 2018;38:6841–53. https://doi.org/10.1523/JNEUROSCI.0399-18.2018.
[33] Keramati M, Durand A, Girardeau P, Gutkin B, Ahmed SH. Cocaine addiction as a homeostatic reinforcement learning disorder. Psychol Rev 2017;124:130–53. https://doi.org/10.1037/rev0000046.
[34] Ahmed SH, Koob GF. Transition to drug addiction: a negative reinforcement model based on an allostatic decrease in reward function. Psychopharmacology 2005;180:473–90. https://doi.org/10.1007/s00213-005-2180-z.
[35] Ahmed SH. Transition from moderate to excessive drug intake: change in hedonic set point. Science 1998;282:298–300. https://doi.org/10.1126/science.282.5387.298.
[36] Keramati M, Ahmed SH, Gutkin BS. Misdeed of the need: towards computational accounts of transition to addiction. Curr Opin Neurobiol 2017;46. https://doi.org/10.1016/j.conb.2017.08.014.
[37] Bellman R. On the theory of dynamic programming. Proc Natl Acad Sci USA 1952;38:716–9.
[38] Friston K, Ao P. Free energy, value, and attractors. Comput Math Methods Med 2012;2012:1–27. https://doi.org/10.1155/2012/937860.
[39] Pezzulo G, Rigoli F, Friston K. Active inference, homeostatic regulation and adaptive behavioural control. Prog Neurobiol 2015:1–19. https://doi.org/10.1016/j.pneurobio.2015.09.001.
[40] Friston KJ , Samothrakis S, Montague PR. Active inference and agency: optimal control without cost functions. Biol Cybern 2012;106(8–9):523–41. https://doi.org/10.1007/s00422-012-0512-8.
[41] Bogacz R. A tutorial on the free-energy framework for modelling perception and learning. J Math Psychol 2015:1–14. https://doi.org/10.1016/j.jmp.2015.11.003.
[42] Friston K. The free-energy principle: a unified brain theory?. Nat Rev Neurosci 2010;11:127–38. https://doi.org/10.1038/nrn2787.
[43] Harper M. Escort evolutionary game theory. Phys D, Nonlinear Phenom 2011;240:1411–5. https://doi.org/10.1016/j.physd.2011.04.008.
[44] Skene K. Life's a gas: a thermodynamic theory of biological evolution. Entropy 2015;17:5522–48. https://doi.org/10.3390/e17085522.
[45] Stroustrup N, Ulmschneider BE, Nash ZM, López-Moyado IF, Apfeld J, Fontana W. The Caenorhabditis elegans Lifespan machine. Nat Methods 2013;10:665–70.
[46] Sommerfield AJ, Deary IJ, Frier BM. Acute hyperglycemia alters mood state and impairs cognitive performance in people with type 2 diabetes. Diabetes Care 2004;27:2335–40. https://doi.org/10.2337/diacare.27.10.2335.
[47] Cryer PE. Hypoglycemia, functional brain failure, and brain death. J Clin Invest 2007;117:868–70. https://doi.org/10.1172/JCI31669.
[48] Keys A. Experimental studies of starvation on men. Bull Chic Med Soc 1946;49:42–6.
[49] Osier FHA. Abnormal blood glucose concentrations on admission to a rural Kenyan district hospital: prevalence and outcome. Arch Dis Child 2003;88:621–5. https://doi.org/10.1136/adc.88.7.621.
[50] Wei M, Gibbons LW, Mitchell TL, Kampert JB, Stern MP, Low Blair SN. Fasting plasma glucose level as a predictor of cardiovascular disease and all-cause mortality. Circulation 2000;101:2047–52. https://doi.org/10.1161/01.CIR.101.17.2047.
[51] Brett JR. Some lethal temperature relations of algonquin park fishes. The University of Toronto Press; 1944.
[52] Stelfox HT, Ahmed SB, Khandwala F, Zygun D, Shahpori R, Laupland K. The epidemiology of intensive care unit-acquired hyponatraemia and hypernatraemia in medical-surgical intensive care units. Crit Care 2008;12:R162. https://doi.org/10.1186/cc7162.