nature machine intelligence

Perspective

https://doi.org/10.1038/s42256-026-01296-8

Life-inspired interoceptive artificial intelligence for autonomous and adaptive agents

Received: 17 March 2025

Accepted: 29 July 2026

Published online: 26 August 2026

Check for updates

Sungwoo Lee[1,2,3,4,9], Younghyun Oh[1,2,3,4,9], Hyunhoe An[1,2,3,4], Hyebhin Yoon[1,2,3,4,5], Karl J. Friston[6], Seok Jun Hong[1,2,3,4,7,8] ✉ & Choong-Wan Woo[1,2,3,4,7] ✉

Despite remarkable advances in artificial intelligence (AI), building agents that can autonomously pursue goals while adapting to continuously changing environments remains a fundamental challenge. Living organisms naturally exhibit such adaptive autonomy, offering a compelling source of inspiration for how artificial systems might achieve similar capabilities. Here we focus on interoception—the process of monitoring and regulating internal bodily states to maintain internal homeostasis—which underwrites an organism's survival. Developing interoceptive AI requires abstracting internal states from their biological instantiation into functional and mathematical representations that are applicable to artificial systems. This requires explicit factorization of state variables representing internal and external environments, together with mathematical formalization of life-inspired properties governing internal-state dynamics. Importantly, internal states can also function as universally available and intrinsically valuable contexts, serving as stable reference signals that modulate learning and behaviour under changing external environments. In addition, interoception-related biological processes, such as neuromodulatory mechanisms, can be incorporated to support context-dependent adaptive behaviour. Overall, this Perspective presents a unified account on how interoception can enhance autonomy and adaptivity in artificial agents by integrating insights from cybernetics with recent advances in theories of life, reinforcement learning and neuroscience.

Although notable advances in artificial intelligence (AI) have led to remarkable achievements in many domains, such as language models, arcade games, visual processing and quadrupedal robot walking[1–4], these successes have largely been achieved under task-specific, externally defined objectives and relatively stable conditions. However, building autonomous and adaptive agents remains an open challenge. To be autonomous, agents must make decisions and execute actions depending on their own goals. To be adaptive, agents must be able to dynamically reconfigure their state representations in response to changing environments and adjust their actions accordingly. Conventional artificial agents, however, rely on pre-engineered inputs to establish new goals[5] and are often incapable of adapting to environmental changes[6]. These challenges are particularly important, as many industrial fields increasingly demand more capable agents for complex tasks such as robot exploration and rescue tasks in space. In these tasks, robots must manage system health and complete missions

A full list of affiliations appears at the end of the paper. ✉e-mail: hongseokjun@skku.edu; waniwoo@skku.edu

https://doi.org/10.1038/s42256-026-01296-8

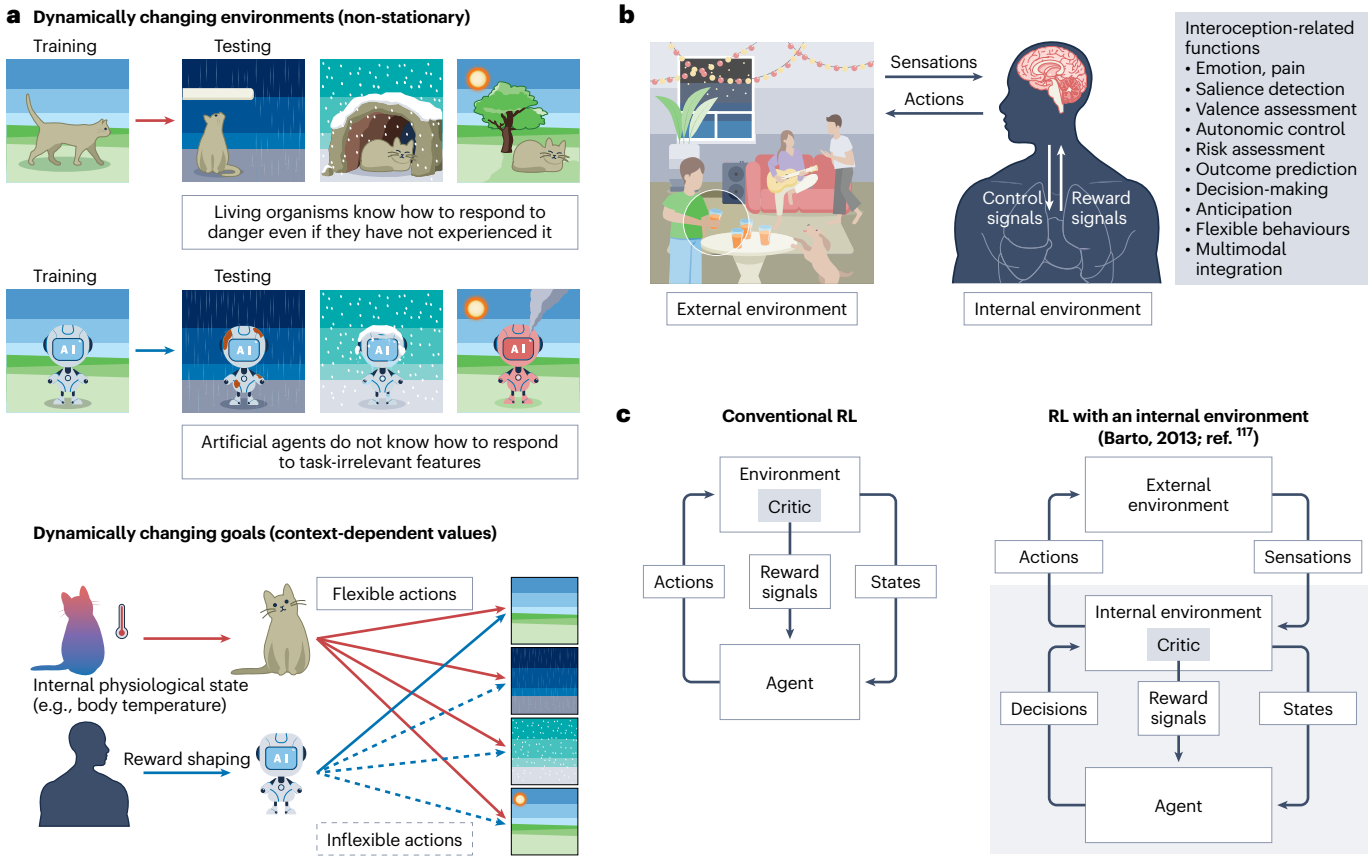

Fig. 1 | Interoception for autonomous and adaptive agents. a, Challenges in training artificial agents. Examples of dynamically changing environments (top) and goals (bottom), which are the two main targets of the interoceptive AI framework. These are substantial challenges in current AI systems. Top (adaptivity): animals can adapt to novel and changing environments by utilizing stable internal values (for example, the cost of being too hot, cold or wet) without explicit training. Unlike animals, however, robots have no internal values and thus find it difficult to adapt to changing and new environments. Bottom (autonomy): animals can easily perform flexible actions with dynamically changing goals, but robots need a designer's additional instructions to change their goals. **b**, Interoception in the brain. Interoception has a crucial role in providing a stable and context-dependent reference for humans and animals. For example, reward signals are generated from the internal environment (upwards arrow) and report an organism's homeostasis (for example, thirst). Interoception then serves to contextualize exteroception in the brain, making some stimuli (for example, beverages) more salient than others to achieve homeostasis (for example, amount of water in the body). The brain can also send a descending control and regulatory signal (downwards arrow) to the body (for example, allostasis)[49,116]. Interoception is known to have an essential role in diverse cognitive and affective functions and their interactions. **c**, Relevant RL frameworks. Left: the conventional RL framework without an internal environment state. In conventional RL, rewards stem from the external environment state. Right: a previous study[117] proposed an alternative model to study intrinsically motivated behaviours, which suggests that reward signals should come from the internal environment.

autonomously with minimal human intervention[7,8]. To overcome these challenges, new ideas from other fields need to be integrated into existing AI frameworks, and we can draw inspiration from living organisms. These organisms exhibit inherent autonomy and adaptivity—for instance, even single-cell eukaryotes can adjust their goals based on their needs and varying conditions[9]. Additionally, living organisms possess innate values, reflecting evolutionarily shaped preferences[10], to avoid or escape hostile environments to preserve their well-being. In essence, living agents inherently know how to choose their own goals and adapt their behaviours in response to changing environments, with the primary goal of maintaining homeostasis within their internal environment—a crucial factor for survival (Fig. 1a).

In this Perspective, we propose an interoceptive AI framework that endows an artificial agent with an internal environment and interoceptive inputs, which could provide contextual information that supports both autonomy and adaptivity. This framework aims to provide a generalizable abstraction of internal states and their contextual role for the modulatory mechanisms to build an agent that can maintain homeostasis within its internal environment state, which requires the ability to monitor its internal state (that is, interoception). In developing this framework, we integrate the original concepts from cybernetics and combine them with recent theoretical work on life and reinforcement learning (RL)[11,12] to formalize the internal state as a 'universal and valuable context'—universal because, even when external cues are sparse, ambiguous or rapidly changing, internal states provide a consistently available, task-agnostic source of contextual information, and valuable because they are intrinsically linked to the reward system[11,12]. Furthermore, by incorporating insights from robotics[13–15], AI[16] and neuroscience[17–20], we emphasize the significance of modulatory computations that leverage the internal state as a context, suggesting a promising direction for advancing autonomous and adaptive AI. We argue that this framework holds the potential to address critical challenges in RL research, such as the exploration–exploitation trade-off and the stability–plasticity dilemma, while also serving as a computational framework for interoception and affect[21,22].

The remainder of this Perspective is organized as follows. We first review interoception in both natural and artificial agents, emphasizing its regulatory roles across biological and engineered systems. We then outline the key elements required to implement internal states in artificial agents and formalize them within an extended Markov

decision process (MDP) framework. Building on this formalization, we present one example implementation of interoceptive AI. We subsequently examine neuromodulatory mechanisms as a potential computational link between internal states and autonomous and adaptive behaviour. Finally, we discuss how interoceptive AI can function as a computational model of interoception and affect. We conclude by considering broader implications, including ethical considerations and future directions.

Interoception for natural and artificial agents

Interoception indicates a biological process to monitor the physiological variables constituting internal states (or internal milieu[23]), such as glucose levels, oxygen levels and blood pressure[24]. This internally oriented process is, as the name indicates, a contrasting concept to its counterpart, 'exteroception', a typically known perceptual process to monitor sensory stimuli in the external environment. In neuroscience, interoception has been actively studied in conjunction with other mental faculties such as feelings[25–27], emotions[28–30], cognition[31], psychopathology[32] and consciousness[33], highlighting its pivotal roles across various mental functions of agents (Fig. 1b). While interoception is traditionally defined as afferent sensing, its functionality is operationally part of a closed feedback loop with efferent regulation. This core cybernetic principle, which integrates the interoceptive sensing with control to enable stability and error correction, has also been adopted in diverse engineering fields[11,34]. For instance, in the field of space robotics and aerospace, methods for monitoring the structural integrity of hardware, such as plastic deformation and fatigue damage, or overall system fault detection, have been studied and implemented, yet under different names, such as 'prognostics and health management' or 'integrated system health management'[35,36]. Modern AI technologies, such as deep learning, have recently been integrated with these systems to predict and manage complex system health[37–39]. Building on this principle of internal monitoring, the interoceptive AI framework provides a unifying abstraction that can help extend such capabilities to a broader class of embodied agents. For example, in physical AI systems[40,41], interoceptive representations can support disaster-response robots in regulating internal stress or pressure under extreme conditions, assist household service robots in managing overheating and mechanical wear, and enable logistics agents to track actuator fatigue over prolonged operation.

A critical aspect of interoception lies in its integrative and regulatory functionality[17,42,43]. For example, natural agents (that is, living organisms) continuously monitor their internal physiological states while performing complex tasks and adapt their actions based on their internal states. Significant deviations in variables such as glucose levels or body temperature prompt animals to prioritize restoring these variables to homeostatic ranges. This results in autonomous shifts in goals, such as switching from exploring the environment to foraging for food or seeking a suitable place to regulate body temperature[44–46]. This adaptive mechanism, often described as allostasis, reflects predictive and context-sensitive regulation driven by learned expectations and anticipated future demands[47–49], in contrast to the more reactive nature of homeostasis[50]. Within the brain, such mechanisms depend on coordinated, distributed modulation across multiple brain systems. The neuroanatomical structure of the interoceptive signal pathway, which passes through the brain's modulatory centres, is positioned to influence higher-order association brain areas[17], including the prefrontal cortex, insula and anterior cingulate cortex, regions important for abstracting value representations[51] and multimodal integration[52–56]. These characteristics may provide valuable insights for the development of autonomous and adaptive artificial agents.

Previous research in robotics has introduced artificial neuromodulatory and neuroendocrine systems to enhance autonomy and adaptivity, yet without fully leveraging interoceptive mechanisms in living organisms[57–60]. For instance, it has been demonstrated that neuromodulatory systems can be introduced to induce behavioural changes in robots, allowing them to adapt more effectively to their environment by dynamically modifying behaviour in response to newly incoming information[61]. Similarly, artificial neuroendocrine systems have been employed to adjust parameters, emulating hormonal mechanisms to modulate a robot's sensitivity to stimuli[58]. These hormonal adjustments can regulate tolerances and sensitivities to both internal and external conditions (for example, a robot in an energy-rich environment tolerates larger energy deficits), thereby promoting adaptation to diverse environments. Despite these advancements, however, most studies remain focused on specific tasks or algorithms, lacking an integrative framework that incorporates ideas from neurobiological mechanisms of interoception into the design of artificial agents or robots to foster the development of more autonomous and adaptive agents[7,8].

Basic elements for implementing internal states

The first step for developing interoceptive AI is the implementation of internal states in artificial agents. This can be achieved by adopting a functional interpretation of internal states, drawing on concepts from cybernetics and modern theories of life. Such an approach requires a shift from a spatio-anatomical perspective (that is, internal states confined within a biological body) to a functionalist view, acknowledging that natural and artificial agents can have distinct forms of internal states, with artificial agents operating at an abstract level. The basic functional characteristics of internal states include (1) factorization, (2) stable state dynamics and (3) being a source of primary reward.

The first characteristic, 'factorization' with respect to the external environment, is a fundamental feature of living organisms. For example, internal states in humans, such as body temperature, are 'factorized' or separated from external variables such as ambient temperature. This factorization allows agents to insulate their internal variables effectively, thereby establishing a clear boundary between themselves and their surroundings. By explicitly considering internal states, the interoceptive AI framework endows an agent with an internal milieu that is distinct from its external environment. Specifically, the factorization enforces internal and external states to have their own state dynamics[62]—for instance, the dynamics of body temperature can be distinguished from those of environmental temperature. Importantly, the factorization does not eliminate the interactions between these states. For example, ambient temperature can influence body temperature, but this interaction is mediated by boundary states, such as the skin, allowing for selective influence of external states on internal ones, unlike the direct interactions among external states. This selective interaction contributes to the second characteristic of internal states: relative stability compared to external states. For instance, while the temperature of the external environment may vary throughout the year, the human body consistently maintains a temperature of ~36.5 °C. However, this stability does not arise automatically from factorization; it requires active regulation through negative feedback mechanisms, in a process known as homeostasis[63]. Furthermore, successful regulation requires active engagement with the external environment while seamlessly integrating internal and external information. Lastly, maintaining homeostasis serves as the primary source of reward for living organisms—that is, survival[64–66]. Deviations of internal states beyond their optimal range, known as the viability zone, lead to failure to survive. In this sense, internal states can be viewed as a set of variables that define the survival of an agent.

These characteristics are inspired by living organisms, laying the groundwork for understanding how internal-state variables guide and regulate their internal and external actions in real-life situations. Internal states with these characteristics have the potential to serve as a universal and intrinsically valuable context, meaning that they provide consistent and structured information that is important for survival and remains relatively stable despite external variability. In the

https://doi.org/10.1038/s42256-026-01296-8

BOX 1

Markov decision process

Markov decision processes (MDPs) are a well-established framework for formalizing sequential decision-making problems[87]. They are defined with a set of states $s \in S$ at any given time step, and at each step, an agent selects an action $a \in A$ to maximize the total reward (or minimize cost). In an MDP, the state transitions occur based on a state transition probability, which defines the environment's dynamics. This probability determines how likely a previous state s transitions to the next state s', given that the agent performs an action a : $p(s'|s,a) \doteq P\{S_{t+1} = s'|S_t = s, A_t = a\}$. The action results in a reward for an agent at each step. The reward is determined by a reward function that assigns a single numeric value, denoted as r, to each state-action pair: $r_t = R(s_t, a_t)$. The objective of an MDP is to select an action (or a sequence of actions) that maximizes the cumulative reward in the long run. Reinforcement learning (RL), a branch of AI research, relies on MDPs to solve various decision-making problems. Many RL algorithms utilize a value function, which represents the total amount of expected reward starting from the initial state s while following a particular policy π (the probability of selecting possible actions given a state): $v_\pi(s) \doteq E_\pi\{\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}|S_t = s\}$, where γ is a discount factor for future rewards. While the components of an MDP, such as states, state transition probabilities and reward functions, can be flexibly configured to meet research objectives, it is essential to adhere to the mathematical constraints of the MDP. For example, the MDP framework is constrained by the Markov property, which posits that each state encapsulates all relevant historical information. In cases where states are not fully observable, the partially observable Markov decision process (POMDP) can be used[119]. The POMDP introduces an observation function (a.k.a. likelihood or emission function) to establish a probabilistic mapping between states and observations, enabling agents to infer states based on observations. Further expanding on POMDP, the mixed observability Markov decision process (MOMDP) incorporates states with varying levels of observability[120]. This adaptability highlights the versatility of the MDP framework, allowing it to accommodate diverse problem settings by introducing different types of mathematical constraint.

Although conventional RL models typically assume stationary MDP, in which state transition probabilities and reward functions remain unchanged over time[121], this stationary assumption is often violated in real-world scenarios[6,122]. Living organisms, for instance, frequently encounter situations where their tasks and goals are continuously changing (that is, non-stationary reward functions), and their environments are also changing (that is, non-stationary state transition probabilities). To address this non-stationarity problem, multiple approaches have been proposed (for a comprehensive review on this topic, see ref. 6), with some using factorized task state variables to represent dynamically changing tasks[122,123]. Our interoceptive AI framework adopts a similar strategy by factorizing the state space into internal and external states, with internal states having greater stability. A unique aspect of our approach lies in the interpretation of stationary components as internal physiological states, which we endow with multiple life-inspired features. Although detailed physiological states are too complex to implement directly, we abstract the shared computational role of internal states—a role supported by theoretical work on autonomy, dynamical independence, nonequilibrium processes and informational boundaries[124–130]. At the same time, we acknowledge that fully characterizing and implementing their detailed biological mechanisms remains an open direction for future research[12,63,69].

case of living organisms, internal physiological states such as glucose levels, body temperature and other physiological states act as reference points for interpreting the environment and determining appropriate actions. By relying on the internal-state information, living organisms effectively predict and calculate rewards and adapt to continuously changing external environments[42].

These ideas have important antecedents in earlier literature, but their further development and exploration have been limited, probably due to their origins in distinct and unconnected fields and the lack of an integrative framework. For example, the idea of establishing a functional and computational link between internal-state variables and survival was first proposed in cybernetics by W. Ross Ashby[11]. He described an organism as a dynamical system and introduced the notion of essential variables—an abstracted form of the internal-state variables that define survival. Examples of essential variables in animals include body temperature, glucose levels and blood pressure, all of which must be maintained within specific bounds to ensure survival[67]. A related idea emerged from the RL field, where Singh et al.[68] proposed to divide state variables into internal and external ones, with internal states providing the reward signal. More recent studies have also explored the concept of homeostasis in the context of RL by incorporating the idea of internal states[63,69]. Our interoceptive AI framework builds on and extends these efforts by treating internal states not merely as sources of reward, but as universal and valuable contextual variables and modulatory signals that shape learning and behaviour. This generalization brings together previously disconnected conceptual threads and is intended to motivate future research into the computational and functional implications of interoceptive principles.

We can achieve this by formalizing the internal states within the MDP framework, which can then be applied to diverse domains such as transportation, manufacturing and financial modelling, providing a novel perspective on the functionality of the internal states beyond physically embodied agents[70] (Box 1). The basic set-up for implementing internal states in the interoceptive MDP framework involves three major steps: (1) factorizing internal states by separating their state transition probabilities from those of external states, (2) applying selective (or sparse) interactions between internal and external state transition probabilities, and (3) mapping the primary reward function onto the dynamics of internal states (Fig. 2). These steps represent the bare minimum for the implementation of internal states, providing a foundational framework upon which researchers can incorporate additional interoception and life-inspired features. This general formulation formalizes and extends existing ideas. For example, homeostatic RL has previously demonstrated the viability of reward mapping onto internal dynamics. However, the explicit factorization of internal states and their placement within non-stationary external environments enables a new interpretation; that is, internal states can serve as universal contextual variables—which can be deployed across diverse situations—that modulate cognition and behaviour. This redefinition, in turn, generates new problem settings that were not furnished by previous frameworks, as elaborated in the following section.

Implementing interoceptive AI and the EVAAA benchmark

One example of instantiating the interoceptive AI framework comes from the recently proposed EVAAA (Essential Variables

https://doi.org/10.1038/s42256-026-01296-8

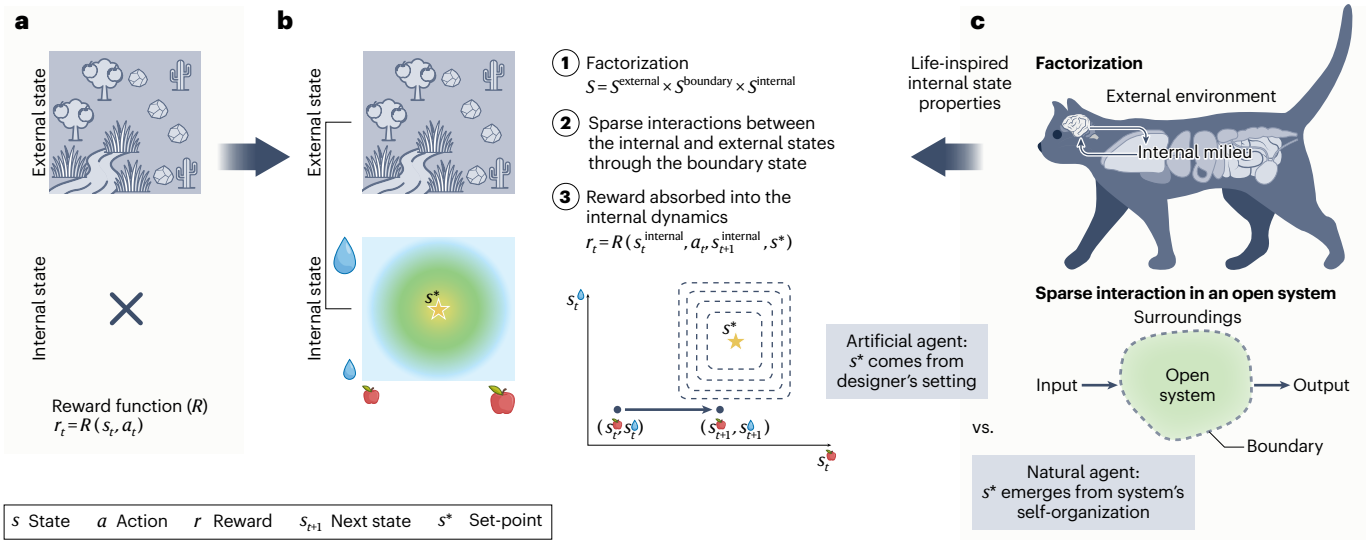

Fig. 2 | Basic elements for implementing internal states. This figure provides an example of interoceptive AI in MDPs. **a**, Conventional RL. Such models only consider external states. For example, Sutton and Barto's RL textbook says that "anything that cannot be changed arbitrarily by the agent is considered to be outside of it"[87]. In this framework, the reward function is mapped onto the external states. **b**, Interoceptive AI. The basic set-up for implementing internal states can be summarized as follows. (1) Interoceptive AI introduces the internal states into the model through state factorization. (2) Interoceptive AI formalizes stability by structuring internal states to interact selectively or sparsely with external states through boundary states. (3) The reward originates from the internal states: reward is absorbed into the internal-state dynamics because the reward can be expressed in terms of some measure of distance from an internal set point (that is, attracting sets of internal states). **c**, In living organisms, the set point (or set points) of the internal states underwrites self-organization and emerges through natural learning and selection[118]. By contrast, interoceptive AI remains artificial because its set point can be determined by a designer.

in Autonomous and Adaptive Agents) benchmark[71]. EVAAA is an open-source three-dimensional (3D) survival simulation platform designed to train and evaluate embodied, egocentric agents that incorporate the core principles of interoceptive AI (Fig. 3). It implements the proposed MDP formulation, translating abstract principles into concrete, experimentally testable tasks. State factorization is explicitly implemented in EVAAA. The total state space is divided into (1) an external state, encompassing non-stationary external environmental conditions—such as changing obstacle types and positions, changing field temperature, and day–night cycles—captured through rich multimodal sensory inputs (for example, egocentric vision, olfaction, thermoception); and (2) an internal state, defined by multidimensional essential variables such as satiation, hydration, body temperature and damage. These essential variables represent the minimal physiological requirements for survival and provide a universal contextual reference for adaptation across diverse environments.

The benchmark consists of two complementary domains: (1) a progressive survival curriculum, in which agents learn to maintain multiple essential variables across increasingly complex, dynamic environments; and (2) a suite of unseen experimental testbeds, inspired by animal behavioural experiments, including the 'two-resource choice task', where agents must select among competing resources based on internal needs, and the 'risk-taking task', where agents weigh potential bodily harm against resource acquisition. These testbeds are designed to probe generalization under novel external perturbations. This design operationalizes two central principles of interoceptive AI. First, homeostatic regulation serves as the primary organizing mechanism. A single homeostatic reward function applies universally across both training and testing phases to guide behaviours, requiring no task-specific redefinition. This differs from conventional RL frameworks, which require prespecified rewards. Second, internal states serve as a universal context: beyond their role in reward computation, they provide stable reference signals that ground adaptive behaviour even in unfamiliar or non-stationary environments, enabling a unified control architecture to generalize across tasks and conditions.

By preserving a stable interoceptive context across changing environments, the benchmark creates the ideal conditions to test how robust modulatory mechanisms could enable learning and adaptation under volatile conditions. Its MDP-based, algorithm-agnostic formulation enables systematic comparisons across different classes of algorithm. Model-free RL allows agents to reactively correct deviations in internal states, reflecting homeostatic control[50]. When augmented with model-based or predictive components, agents can anticipate future deviations and proactively adjust internal set points or behaviours—an artificial analogue of allostatic regulation[47–49]. The survival-based metrics further enable researchers to assess which algorithms produce more adaptive behaviour and better leverage internal-state context to influence learning, memory, motivation and decision-making, thereby linking theoretical insights with empirical evaluation in interoceptive AI.

We note that the EVAAA benchmark, and the interoceptive AI framework more broadly, are intended as foundational and exploratory abstractions, rather than high-fidelity models of biological mechanisms of interoception. The full representational diversity, sparse interaction, and dynamics of interoceptive signals, as well as their multisensory integration, remain in the early stages of understanding in biology. Accordingly, our current framework does not attempt to replicate the full biological details, but instead abstracts the shared computational roles of interoceptive signals as contextual modulators of learning and control. EVAAA should therefore be viewed as a scaffold for future work, enabling systematic investigation of internal-state representations and modulatory mechanisms under controlled conditions. Future extensions may incorporate richer, hierarchical and multiscale implementations that more closely approximate biological complexity. In addition, by operating in a simulation-first setting, EVAAA also provides a principled environment in which the potential risks—such as refusal-like behaviours, misaligned priorities or over-prioritization of internal needs—can be systematically probed through controlled perturbations of internal variables while task goals are held fixed, before such implementations are considered for integration into more advanced models or real-world robotic systems.

https://doi.org/10.1038/s42256-026-01296-8

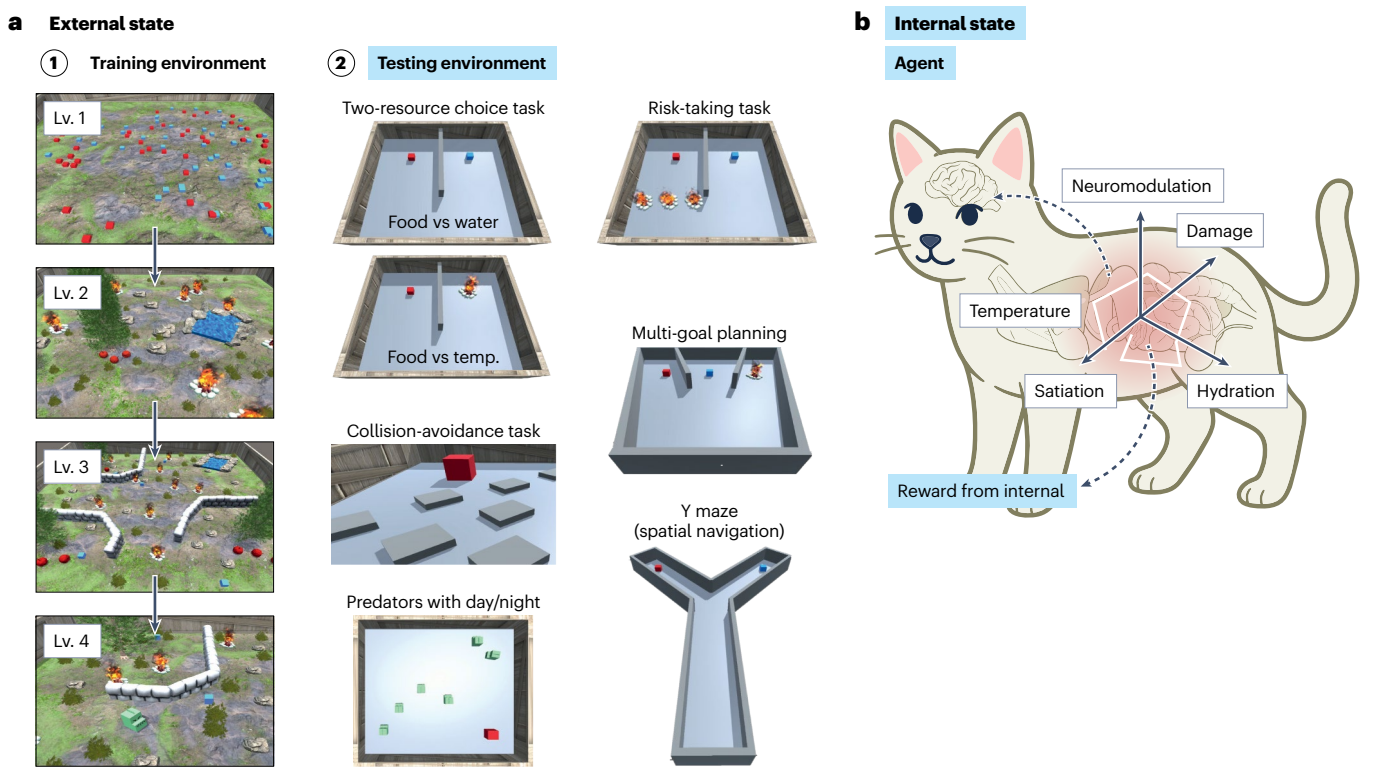

Fig. 3 | Overview of the EVAAA benchmark for implementing interoceptive AI. This figure illustrates the integration of external non-stationary environments and internal-state regulation in a virtual environment platform for Essential Variables in Autonomous and Adaptive Agents (EVAAA). **a**, EVAAA adopts a two-tiered design consisting of a naturalistic training curriculum with progressively increasing complexity (levels 1–4) and a set of controlled experimental testbeds that are unseen during training and used exclusively for evaluation. Training environments (left) support survival-oriented behaviours such as resource foraging, obstacle navigation, predator avoidance, and adaptation to day–night cycles. Experimental testbeds (right) isolate specific cognitive and behavioural tests—including two-resource choice, risk-taking, collision avoidance, multi-goal planning and spatial navigation (Y maze)—to assess autonomy and adaptivity under internal-state constraints. This separation between training and testing environments introduces non-stationarity and distributional shift, highlighting the role of internal states as a source of task-agnostic contextual information. Across all environments, red cubes represent food resources, and blue cubes represent water resources. **b**, The EVAAA agent maintains multiple essential internal variables (for example, satiation, hydration, temperature and damage), which evolve through interaction with the environment and are continuously monitored via interoceptive signals. These internal states generate intrinsic reward signals and provide contextual input that can be used to implement modulatory computations, shaping learning dynamics, exploration–exploitation balance, and policy adaptation. Blue shaded boxes indicate features emphasized in the EVAAA framework.

Neuromodulatory computations for interoceptive AI

Neuromodulatory functions serve as a key mechanism that bridges internal states with autonomous and adaptive behaviours, as previous neuroscience literature has highlighted the role of a neuromodulatory process in mediating the interactions between the body and the brain to maintain homeostasis and facilitate adaptive responses to dynamically changing environments[17,20,42]. In *Drosophila*, for example, several studies have shown that internal states such as hunger can modulate sensory processing through neuromodulatory mechanisms, influencing foraging behaviour. During starvation, short neuropeptide F enhances sensitivity to food-related odours via the DM1 glomerulus, while tachykinin suppresses aversive responses in the DM5 glomerulus, enabling adaptive foraging[72]. In addition, beyond sensory-driven behaviours, internal states can dynamically shape the balance between exploration and exploitation, with nutrient-deprived *Drosophila* tending to exploit specific food patches while satiated flies engage in broader exploration, thereby optimizing future foraging opportunities[73,74]. Moreover, metabolic cues influence memory formation, as hunger modulates both the encoding and retrieval of learned food associations through neuromodulatory circuits in the mushroom bodies[75–77].

To implement modulatory mechanisms in artificial systems, it is essential to abstract their computational functionalities, which primarily involve modulating neuronal sensitivity to inputs[78]. Within the proposed interoceptive AI framework, both neuromodulators and hormones can be computationally unified under the broader notion of neuromodulation[42]. In this view, both serve as mechanisms that translate internal states into adaptive control signals, differing only in their temporal and spatial scales of influence, with neuromodulators acting relatively locally and rapidly and hormones acting relatively globally and more slowly[79]. These computations ensure the efficient encoding of sensory information through two key mechanisms: multiplicative gain modulation, which amplifies relevant signals to enhance signal-to-noise ratios, and additive gain modulation, which adjusts neural excitability to maintain a balance between responsiveness to new inputs and the stability of ongoing processes. Combined with interoceptive inputs—proposed as universal and valuable contexts—these mechanisms could have a crucial role in enabling context-dependent adaptive responses by dynamically regulating the system's functionalities[80].

The neuromodulatory mechanisms have been applied in robotics[60,81] and RL[82,83] to design adaptive robots and AI agents. For instance, neuromodulators such as dopamine, serotonin, acetylcholine and noradrenaline have often been formalized as hyperparameters, with neuromodulation conceptualized as the tuning of hyperparameters, including reward sensitivity, exploration–exploitation balance, and learning rates[60,82]. In addition, a Bayesian interpretation of neuromodulatory computation suggests that neuromodulators

https://doi.org/10.1038/s42256-026-01296-8

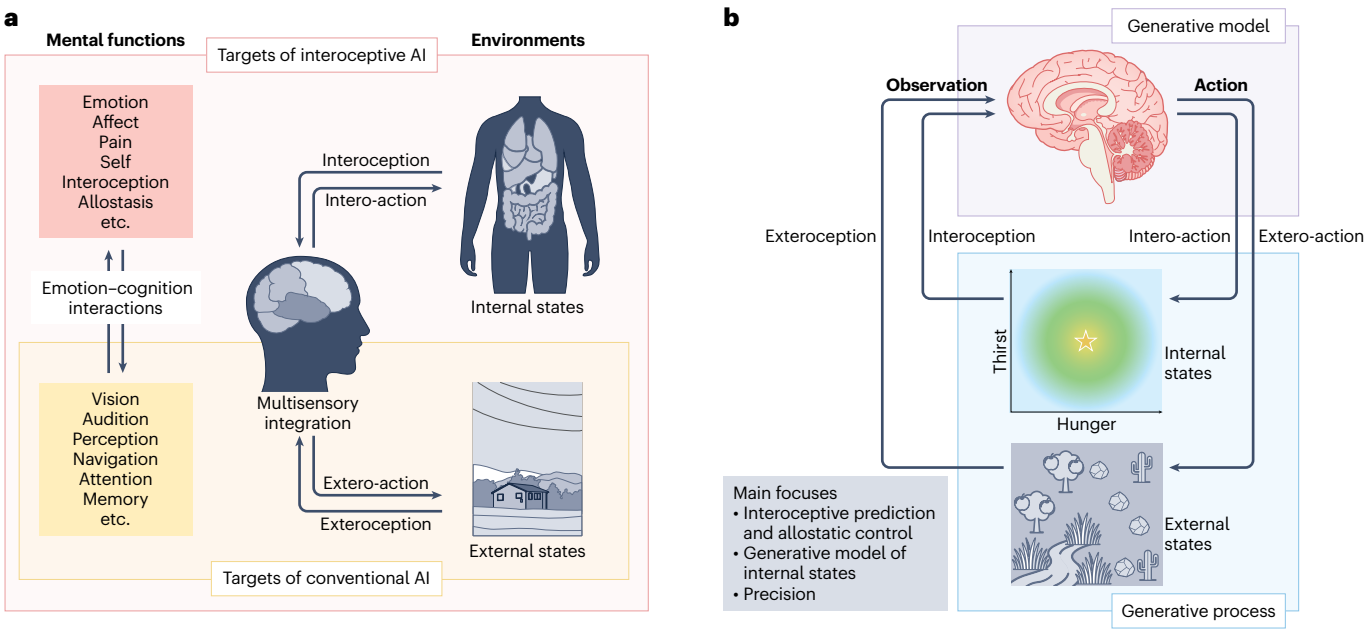

Fig. 4 | The interoceptive AI framework for affective neuroscience. **a**, Interoceptive AI as a computational model of interoception and affect. Conventional AI models have provided a useful framework for the study of perceptual and cognitive functions, such as vision, navigation and memory (the yellow area). They usually consider only the mental functions and processes pertaining to extra-personal environments (for example, exteroception), and thus they are not suitable for studying internal environment-related functions, such as emotion, pain and interoception (the red area). Alternatively, the interoceptive AI framework encompasses an internal environment, providing a computational model for interoception-related functions and their interactions with exteroception-related functions. **b**, Interoceptive active inference: a computational model of interoception. According to interoceptive active inference, an internal environment state is considered hidden from the brain. The brain only has access to the observations that the internal and external states generate (the generative process), and, based on the observations, the brain forms an internal model of how the observations are generated (the generative model). In interoceptive active inference, the brain is assumed to engage autonomic reflexes to realize homeostatic set points for attracting sets. The distance between interoceptive inputs and set points is measured by the surprisal (or prediction error) in the same way that exteroceptive predictive coding uses surprisal or prediction errors for perception and proprioceptive prediction errors for motor function. The main focuses of interoceptive active inference include belief updating, required to generate appropriate (interoceptive) set points in the form of predictions and allostatic control and contextualizing the precision of accompanying prediction errors.

encode uncertainty or its complement, precision[20]. These concepts have recently been extended to deep neural networks, incorporating a primary network for task processing and an auxiliary neuromodulatory network that dynamically adjusts the primary network's weights and activations based on contextual inputs. This architecture has been shown to mitigate catastrophic forgetting by regulating synaptic plasticity, enabling the system to preserve previously learned knowledge while accommodating new information via the selection of sub-networks in a context-dependent manner[84–86]. Building on these developments, interoceptive AI proposes using internal states as a universal source of contextual signals for modulatory control under a non-stationary environment.

Interoceptive AI incorporating modulatory mechanisms has the potential to address several long-standing dilemmas in AI. One such dilemma is the exploration–exploitation trade-off, in which agents risk missing potentially better outcomes by adhering to known actions (that is, exploitation) or failing to utilize known actions effectively due to excessive exploration[87]. The common approach in RL to address this dilemma involves using ad hoc novelty-based intrinsic rewards, encouraging agents to explore more when encountering novel situations in their environment[88–91]. However, in non-stationary environments, this method can result in persistent exploration, as agents would continuously encounter novel situations[90,91].

The interoceptive AI framework offers a life-inspired, internal-state-driven alternative. As the animal literature shows, the decision to explore or exploit is often driven by internal needs[42,73]: a hungry animal exploits its existing knowledge to locate food, whereas a satiated animal explores new territory. EVAAA offers one concrete instantiation of how this principle may be implemented in practice. For example, an agent with low satiation (that is, hungry) must learn an exploitative policy to take the fastest path to food. In contrast, an agent with high satiation can afford an exploratory policy to search unknown areas to find other resources. This provides a concrete testbed to quantify and mechanistically test the neuromodulatory control of the agent's policy based on the internal state.

Another dilemma is the stability–plasticity trade-off, a long-standing challenge in machine learning[92] and RL research[6]. It arises when agents must balance retaining prior knowledge (stability) with incorporating new information (plasticity) in non-stationary environments. Current AI systems often suffer from catastrophic forgetting, where new learning overwrites previously acquired knowledge. In the conventional RL framework, agents operating in volatile environments struggle to determine which information to retain or update due to a lack of predefined assumptions regarding environmental stability.

The interoceptive AI framework has the potential to address this dilemma by leveraging the stability of internal states, even in the presence of a non-stationary external environment. EVAAA could be a concrete example to implement this. In EVAAA, the external environment is highly non-stationary (for example, day–night cycles modulate temperature, predators appear, and resources are randomly generated or change their appearance). This confronts the agent with a constant stability–plasticity problem: it must learn new information (plasticity) without overwriting old, critical knowledge (stability). An interoceptive agent can use its stable internal state as an anchor. For example, a neuromodulation mechanism could use the internal state to select different sub-networks in neural networks for different contexts

(for example, a 'daytime/safe' network versus a 'nighttime/danger' network)[84], allowing it to adapt while preserving core knowledge.

Interoceptive AI as a computational model of interoception and affect

The interoceptive AI framework could serve as a computational model for interoception and affect (Fig. 4a). With remarkable advances in AI, neuroscientists have begun to use AI agents as computational models of cognitive functions[93]. By comparing the AI agents' behaviours and neural representations with those of humans or other animals, one can study the underlying neurocognitive principles that are often difficult to investigate using traditional experimental methods in neuroscience. However, implementing this approach requires well-defined tasks for AI. Compared to cognitive and decision-making processes, interoception and affect-related processes currently have limited tasks for testing. Recently, researchers have proposed that homeostasis and allostasis can offer key testable bases to study affective neuroscience[22]. Also, other researchers have emphasized the importance of distinguishing different types of reward in understanding emotional experiences: one originating from bodily built-in value systems, referred to as proxy rewards or shaping, and the other originating from internal states, that is, primary rewards[10,65,94]. These approaches, along with foundational sub-cortical accounts[27], reflexive/pre-reflexive mechanisms, and standard learning models (for example, Pavlovian, habitual), provide a rich theoretical backbone for computational modelling. Furthermore, the relationship between emotions and homeostasis has been studied in the context of robotics[13–15,95]. Building on all this prior knowledge, interoceptive AI is intended as a general computational formalism that can be used to study key topics in affective neuroscience.

Previous studies have proposed computational models of interoception using different algorithmic solutions, such as RL, distinct interoceptive predictive coding accounts[21,96], and active inference[30], all of which can be integrated into the interoceptive AI framework. Among these, active inference is one such approach that shares many of the same commitments (Fig. 4b). Although interoceptive AI focuses more on formalizing the internal states for their contextual roles, and active inference emphasizes the belief-updating or inference processes associated with self-organized behaviour, both rest upon the constraints imposed by internal states or interoceptive modalities. The primary question addressed by active inference is 'How do living organisms persist while engaging in adaptive exchanges with their environment?'[97] In many respects, answering this question amounts to formulating the fundamental concepts of interoceptive AI. In addition, one of the unique contributions of active inference to the conventional RL framework is to replace the reward function with prior preferences[97], which, in the context of interoceptive AI, can be interpreted as internal set points or attracting sets of internal states. Active inference has also been compared with conventional RL approaches[98–100], suggesting its ability to effectively handle dynamically changing goals[101] and non-stationary external environments[102], again the two primary objectives that the interoceptive AI framework pursues. Recently, active inference has been applied to interoception, leading to the development of interoceptive active inference[30,103], which provides a computational account of interoception and affect[21,22,104].

Expanding on this foundation, precision weighting is emerging as a key mechanism in active inference that is particularly relevant to interoceptive AI, as it governs the extent to which prediction errors influence belief updating[20]. In active inference, prediction errors are signals that indicate discrepancies between expected and actual sensory inputs. However, not all prediction errors are equally reliable; some carry more meaningful information than others[105]. This selective emphasis on certain prediction errors over others allows for more flexible and adaptive behaviour, as the brain can prioritize important signals while filtering out noise. This precision-weighting mechanism is intricately linked to neuromodulation, for example by dopamine, acetylcholine, noradrenaline, and serotonin, which regulate the precision of sensory signals, including interoception[20]. In active inference, precision modulation serves as a principled mechanism that integrates two distinct computational mechanisms—neural gain control and hyperparameter tuning. Neural gain control dynamically regulates the excitability of neural circuits, determining how strongly prediction errors influence belief updates. Simultaneously, precision modulation is akin to adjusting the generative model's parameters, allowing the system to adapt flexibly to non-stationary environments, much like hyperparameter tuning in RL. By incorporating these principles from active inference, the interoceptive AI framework seeks to provide a formalized model of interoceptive processes and affective states, advancing our understanding of their computational and neural underpinnings.

Ethical implications of interoceptive AI

Introducing internal regulatory variables into artificial agents raises several lines of ethical enquiry that merit continued exploration. If an agent derives intrinsic reward from maintaining its internal variables, could these homeostatic objectives sometimes compete with user-specified missions? It is also reasonable to ask whether interoceptive mechanisms might eventually interact with how artificial agents represent or respond to human welfare. A further concern arises from long-standing associations between embodiment, interoception and conscious experience in biological organisms: might internal state representations, even in simplified AI systems, edge artificial agents towards properties that observers interpret as subjective feelings? These questions highlight areas where the field does not yet have consensus and where additional theoretical and empirical work will probably be required.

Our framework offers one potential avenue for engaging with these issues. Homeostatic variables, although central to the agent's learning dynamics, need not operate as independent survival goals. They can instead be embedded within a hierarchical objective structure in which user-defined tasks remain primary, and internal regulation acts as a constraint that enables goal pursuit rather than competing with it. This organization aligns with many autonomous systems currently operating: a planetary rover cannot perform geological sampling when battery temperatures exceed safe limits; a drone may interrupt a mapping mission if its rotors overheat; and a service robot may throttle its motors or seek charging when experiencing internal strain or low energy. In these examples, maintaining 'bodily health' is instrumental to mission feasibility, not a competing objective. Whether similar hierarchical relationships will continue to hold as agents become more adaptive and open-ended remains an open question that warrants further investigation.

The question of alignment with human values also invites broader reflection. One perspective, motivated by work in social neuroscience and computational models of artificial empathy, suggests that prosocial tendencies may require self-maintenance mechanisms that allow an agent to represent or approximate the internal states of others[106,107]. Biological and social examples, such as caregiving, cooperative foraging or helping behaviours, have been studied within this framework[108–110]. In one recent study it was proposed that prosocial behaviour in artificial agents may require some form of coupling or inference over other agents' homeostatic variables, as explored in multi-agent RL work on homeostatic empathy[111]. These ideas remain at an early stage, but they raise intriguing possibilities for how artificial systems might eventually incorporate socially grounded motivations. In this context, it is relevant that the current interoceptive AI framework models only self-directed internal variables and does not include mechanisms for social inference, affective coupling, or sensitivity to the internal states of humans or other artificial agents. Should future research explore such directions, it would probably introduce a separate set of ethical and conceptual considerations regarding how artificial agents represent, prioritize and respond to the states of others.

Finally, there is an emerging concern that as artificial agents become more embodied and capable, observers may increasingly ascribe consciousness or subjective experience to them based on functional resemblance rather than on genuine computational foundations[112]. Such interpretations could arise even when a system lacks the architectural features that current theories propose as prerequisites for conscious experience, such as global broadcasting, higher-order representations or metacognitive self-models[113]. In our context, interoception and embodiment contribute to adaptive functioning, but they remain insufficient for consciousness on their own. Many autonomous robots already manage internal variables such as battery levels, actuator strain or thermal limits to maintain operational health, yet these forms of bodily regulation carry no implications of subjective awareness[8,37,38]. Clarifying these distinctions may help temper premature interpretations, but the prospect of increasingly sophisticated interoceptive systems raises substantive ethical questions about how such agents might be perceived and evaluated, underscoring the importance of sustained scrutiny and broader interdisciplinary discussion moving forward.

Concluding remarks

Here, drawing inspiration from living organisms, we have introduced the interoceptive AI framework, crafted to enhance the autonomy and adaptivity of artificial agents through the incorporation of an internal environment into the traditional AI framework. The proposed system enables an agent to monitor its internal state and adaptively recalibrate its goals and responses to cope with environmental changes. According to Claude Bernard[114], "the stability of the internal environment is the condition for the free and independent life", and Singh et al.[115] noted that "all rewards are internal". We believe that our life-inspired ideas of state factorization, mapping rewards onto the internal state dynamics, and integrating neuromodulatory mechanisms will serve as essential building blocks for building autonomous and adaptive intelligence. More importantly, our aspiration is that our interoceptive AI framework will deepen our understanding of animal and human intelligence.

Notably, the capabilities enabled by interoception in AI systems could potentially raise ethical and moral considerations. Recognizing these concerns, we emphasize the importance of fostering careful discussions and building a consensus aligned with human values on ethical and governance issues with the development and application of interoceptive AI systems. Such efforts are essential to ensure that these technologies are developed and utilized in ways that uphold societal values and benefit humanity.

References

1. Brown, T. et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems* Vol. 33, 1877–1901 (Curran Associates, 2020).
2. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
3. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 10684–10695 (IEEE, 2022).
4. Miki, T. et al. Learning robust perceptive locomotion for quadrupedal robots in the wild. *Sci. Robot.* **7**, eabk2822 (2022).
5. Colas, C., Karch, T., Sigaud, O. & Oudeyer, P.-Y. Autotelic agents with intrinsically motivated goal-conditioned reinforcement learning: a short survey. *J. Artif. Intell. Res.* **74**, 1159–1199 (2022).
6. Khetarpal, K., Riemer, M., Rish, I. & Precup, D. Towards continual reinforcement learning: a review and perspectives. *J. Artif. Intell. Res.* **75**, 1401–1476 (2022).
7. Gao, Y. & Chien, S. Review on space robotics: Toward top-level science through space exploration. *Sci. Robot.* **2**, eaan5074 https://doi.org/10.1126/scirobotics.aan5074 (2017).
8. Ranasinghe, K. et al. Advances in Integrated System Health Management for mission-essential and safety-critical aerospace applications. *Prog. Aerosp. Sci.* **128**, 100758 (2022).
9. Dexter, J. P., Prabakaran, S. & Gunawardena, J. A complex hierarchy of avoidance behaviors in a single-cell eukaryote. *Curr. Biol.* **29**, 4323–4329.e2 (2019).
10. Dayan, P. "Liking" as an early and editable draft of long-run affective value. *PLoS Biol.* **20**, e3001476 (2022).
11. Ashby, W. R. *Design for a Brain: The Origin of Adaptive Behaviour* (Chapman & Hall, 1960).
12. Friston, K. et al. The free energy principle made simpler but not too simple. *Phys. Rep.* **1024**, 1–29 (2023).
13. Chiba, A. A. & Krichmar, J. L. Neurobiologically inspired self-monitoring systems. *Proc. IEEE Inst. Electr. Electron. Eng.* **108**, 976–986 (2020).
14. Canamero, L. Embodied robot models for interdisciplinary emotion research. *IEEE Trans. Affect. Comput.* **12**, 340–351 (2021).
15. Man, K. & Damasio, A. Homeostasis and soft robotics in the design of feeling machines. *Nat. Mach. Intell.* **1**, 446–452 (2019).
16. Mei, J., Muller, E. & Ramaswamy, S. Informing deep neural networks by multiscale principles of neuromodulatory systems. *Trends Neurosci.* **45**, 237–250 (2022).
17. Teckentrup, V. & Kroemer, N. B. Mechanisms for survival: vagal control of goal-directed behavior. *Trends Cogn. Sci.* **28**, 237–251 (2024).
18. Azzalini, D., Rebollo, I. & Tallon-Baudry, C. Visceral signals shape brain dynamics and cognition. *Trends Cogn. Sci.* **23**, 488–509 (2019).
19. Shine, J. M. Neuromodulatory control of complex adaptive dynamics in the brain. *Interface Focus* **13**, 20220079 (2023).
20. Friston, K. Computational psychiatry: from synapses to sentience. *Mol. Psychiatry* **28**, 256–268 (2023).
21. Petzschner, F. H., Garfinkel, S. N., Paulus, M. P., Koch, C. & Khalsa, S. S. Computational models of interoception and body regulation. *Trends Neurosci.* **44**, 63–76 (2021).
22. Schiller, D. et al. The human affectome. *Neurosci. Biobehav. Rev.* **158**, 105450 (2024).
23. Gross, C. G. Claude Bernard and the constancy of the internal environment. *Neuroscientist* **4**, 380–385 (1998).
24. Engelen, T., Solcà, M. & Tallon-Baudry, C. Interoceptive rhythms in the brain. *Nat. Neurosci.* **26**, 1670–1684 (2023).
25. Craig, A. D. B. How do you feel-now? The anterior insula and human awareness. *Nat. Rev. Neurosci.* **10**, 59–70 (2009).
26. Damasio, A. *The Strange Order of Things: Life, Feeling and the Making of Cultures* (Vintage, 2019).
27. Damasio, A. & Carvalho, G. B. The nature of feelings: evolutionary and neurobiological origins. *Nat. Rev. Neurosci.* **14**, 143–152 (2013).
28. Barrett, L. F. The theory of constructed emotion: an active inference account of interoception and categorization. *Soc. Cogn. Affect. Neurosci.* **12**, 1–23 (2017).
29. Critchley, H. D. & Garfinkel, S. N. Interoception and emotion. *Curr. Opin. Psychol.* **17**, 7–14 (2017).
30. Seth, A. K. & Friston, K. J. Active interoceptive inference and the emotional brain. *Phil. Trans. R. Soc. B* **371**, 20160007 https://doi.org/10.1098/rstb.2016.0007 (2016).
31. Tsakiris, M. & Critchley, H. Interoception beyond homeostasis: affect, cognition and mental health. *Phil. Trans. R. Soc. B* **371**, 20160002 https://doi.org/10.1098/rstb.2016.0002 (2016).
32. Paulus, M. P., Feinstein, J. S. & Khalsa, S. S. An active inference approach to interoceptive psychopathology. *Annu. Rev. Clin. Psychol.* **15**, 97–122 (2019).
33. Seth, A. K. & Tsakiris, M. Being a beast machine: the somatic basis of selfhood. *Trends Cogn. Sci.* **22**, 969–981 (2018).
34. Wiener, N. *Cybernetics or Control and Communication in the Animal and the Machine* (Wiley, 1948).

35. Xu, J. & Xu, L. *Integrated System Health Management*: *Perspectives on Systems Engineering Techniques* (Academic, 2017).
36. Tsui, K. L., Chen, N., Zhou, Q., Hai, Y. & Wang, W. Prognostics and health management: a review on data driven approaches. *Math. Problems Eng.* **2015**, 793161 (2015).
37. Ezhilarasu, C. M., Skaf, Z. & Jennions, I. K. The application of reasoning to aerospace Integrated Vehicle Health Management (IVHM): challenges and opportunities. *Prog. Aerosp. Sci.* **105**, 60–73 (2019).
38. Ellefsen, A. L., Æsøy, V., Ushakov, S. & Zhang, H. A comprehensive survey of prognostics and health management based on deep learning for autonomous ships. *IEEE Trans. Reliab.* **68**, 720–740 (2019).
39. Lee, J. et al. Intelligent maintenance systems and predictive manufacturing. *J. Manuf. Sci. Eng.* **142**, 110805 (2020).
40. Liu, Y. et al. Aligning cyber space with physical world: a comprehensive survey on embodied AI. *IEEE/ASME Trans. Mechatron.* **30**, 7253–7274 https://doi.org/10.1109/tmech.2025.3574943 (2025).
41. Xiang, K. et al. Aligning perception, reasoning, modeling and interaction: a survey on physical AI. *IEEE Trans. Pattern Anal. Mach. Intell.* (in the press).
42. Nässel, D. R. & Zandawala, M. Endocrine cybernetics: neuropeptides as molecular switches in behavioural decisions. *Open Biol.* **12**, 220174 (2022).
43. Chrousos, G. P. Stress and disorders of the stress system. *Nat. Rev. Endocrinol.* **5**, 374–381 (2009).
44. Jiménez-Rodríguez, A., Prescott, T. J., Schmidt, R. & Wilson, S. A framework for resolving motivational conflict via attractor dynamics. In *Biomimetic and Biohybrid Systems; Lecture Notes in Computer Science* (eds Vouloutsi, V. et al.) Vol. 12413, 192–203 (Springer, 2020).
45. Rosado, O. G., Amil, A. F., Freire, I. T. & Verschure, P. F. M. J. Drive competition underlies effective allostatic orchestration. *Front. Robot. AI* **9**, 1052998 (2022).
46. Rosado, O. G., Amil, A. F., Freire, I. T., Vinck, M. & Verschure, P. F. M. J. Biomimetic self-regulation in intrinsically motivated robots. In *ALIFE 2025: Ciphers of Life: Proc. Artificial Life Conference* Vol. 37, 49 (MIT Press, 2025).
47. McEwen, B. Allostasis and the epigenetics of brain and body health over the life course: the brain on stress. *JAMA Psychiatry* **74**, 551–552 (2017).
48. Ramsay, D. S. & Woods, S. C. Clarifying the roles of homeostasis and allostasis in physiological regulation. *Psychol. Rev.* **121**, 225–247 (2014).
49. Schulkin, J. & Sterling, P. Allostasis: a brain-centered, predictive mode of physiological regulation. *Trends Neurosci.* **42**, 740–752 (2019).
50. Cannon, W. B. Organization for physiological homeostasis. *Physiol. Rev.* **9**, 399–431 (1929).
51. De Martino, B. & Cortese, A. Goals, usefulness and abstraction in value-based choice. *Trends Cogn. Sci.* **27**, 65–80 https://doi.org/10.1016/j.tics.2022.11.001 (2023).
52. Gogolla, N. The insular cortex. *Curr. Biol.* **27**, R580–R586 (2017).
53. Engelen, T., Buot, A., Grèzes, J. & Tallon-Baudry, C. Whose emotion is it? Perspective matters to understand brain-body interactions in emotions. *Neuroimage* **268**, 119867 (2023).
54. Grivaz, P., Blanke, O. & Serino, A. Common and distinct brain regions processing multisensory bodily signals for peripersonal space and body ownership. *Neuroimage* **147**, 602–618 (2017).
55. Park, H.-D. & Blanke, O. Coupling inner and outer body for self-consciousness. *Trends Cogn. Sci.* **23**, 377–388 (2019).
56. Simmons, W. K. et al. Keeping the body in mind: insula functional organization and functional connectivity integrate interoceptive, exteroceptive, and emotional awareness: functional organization. *Hum. Brain Mapp.* **34**, 2944–2958 (2013).
57. L'Haridon, L. & Cañamero, L. The effects of stress and predation on pain perception in robots. In *Proc. 2023 11th International Conference on Affective Computing and Intelligent Interaction* 1–8 (IEEE, 2023).
58. Lones, J., Lewis, M. & Cañamero, L. A hormone-driven epigenetic mechanism for adaptation in autonomous robots. *IEEE Trans. Cogn. Dev. Syst.* **10**, 445–454 (2018).
59. Krichmar, J. L. & Hwu, T. J. Design principles for neurorobotics. *Front. Neurorobot.* **16**, 882518 (2022).
60. Avery, M. C. & Krichmar, J. L. Neuromodulatory systems and their interactions: a review of models, theories and experiments. *Front. Neural Circuits* **11**, 108 (2017).
61. Krichmar, J. L. A neurorobotic platform to test the influence of neuromodulatory signaling on anxious and curious behavior. *Front. Neurorobot.* **7**, 1 (2013).
62. Friston, K. Life as we know it. *J. R. Soc. Interface* **10**, 20130475 (2013).
63. Keramati, M. & Gutkin, B. Homeostatic reinforcement learning for integrating reward collection and physiological stability. *eLife* **3**, e04811 (2014).
64. Juechems, K. & Summerfield, C. Where does value come from? *Trends Cogn. Sci.* **23**, 836–850 (2019).
65. Weber, L. A., Yee, D. M., Small, D. M. & Petzschner, F. H. The interoceptive origin of reinforcement learning. *Trends Cogn. Sci.* **29**, 840–854 https://doi.org/10.1016/j.tics.2025.05.008 (2025).
66. Yoshida, N., Sprekeler, H. & Gutkin, B. Linking homeostasis to reinforcement learning: internal state control of motivated behavior. *Curr. Opin. Behav. Sci.* **66**, 101611 (2025).
67. Goldstein, D. S. & Kopin, I. J. Homeostatic systems, biocybernetics and autonomic neuroscience. *Auton. Neurosci.* **208**, 15–28 (2017).
68. Chentanez, N., Barto, A. G. & Singh, S. P. Intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems* Vol. 17 (eds Saul, L. K. et al.) 1281–1288 (MIT Press, 2005).
69. Yoshida, N. Homeostatic agent for general environment. *J. Artif. Gen. Intell.* **8**, 1–22 (2017).
70. Boucherie, R. J. & Van Dijk, N. M. *Markov Decision Processes in Practice* (Springer, 2017).
71. Lee, S. et al. EVAAA: a virtual environment platform for essential variables in autonomous and adaptive agents. In *Advances in Neural Information Processing Systems* Vol. 38 (Curran Associates, 2025).
72. Ko, K. I. et al. Starvation promotes concerted modulation of appetitive olfactory behavior via parallel neuromodulatory circuits. *eLife* **4**, e08298 (2015).
73. Corrales-Carvajal, V. M., Faisal, A. A. & Ribeiro, C. Internal states drive nutrient homeostasis by modulating exploration-exploitation trade-off. *eLife* **5**, https://doi.org/10.7554/elife.19920 (2016).
74. Whitehead, S. C., Sahai, S. Y., Stonemetz, J. & Yapici, N. Exploration–exploitation trade-off is regulated by metabolic state and taste value in *Drosophila*. Preprint at *bioRxiv* https://doi.org/10.1101/2024.05.13.594045 (2024).
75. Krashes, M. J. et al. A neural circuit mechanism integrating motivational state with memory expression in *Drosophila*. *Cell* **139**, 416–427 (2009).
76. Senapati, B. et al. A neural mechanism for deprivation state-specific expression of relevant memories in *Drosophila*. *Nat. Neurosci.* **22**, 2029–2039 (2019).
77. Sgammeglia, N. & Sprecher, S. G. Interplay between metabolic energy regulation and memory pathways in *Drosophila*. *Trends Neurosci.* **45**, 539–549 (2022).
78. Shine, J. M. et al. Computational models link cellular mechanisms of neuromodulation to large-scale neural dynamics. *Nat. Neurosci.* **24**, 765–776 (2021).

79. Kim, S. M., Su, C.-Y. & Wang, J. W. Neuromodulation of innate behaviors in *Drosophila*. *Annu. Rev. Neurosci.* **40**, 327–348 (2017).
80. Ferguson, K. A. & Cardin, J. A. Mechanisms underlying gain modulation in the cortex. *Nat. Rev. Neurosci.* **21**, 80–92 (2020).
81. Kudithipudi, D. et al. Biological underpinnings for lifelong learning machines. *Nat. Mach. Intell.* **4**, 196–210 (2022).
82. Doya, K. Metalearning and neuromodulation. *Neural Netw.* **15**, 495–506 (2002).
83. Daw, N. D., Kakade, S. & Dayan, P. Opponent interactions between serotonin and dopamine. *Neural Netw.* **15**, 603–616 (2002).
84. Velez, R. & Clune, J. Diffusion-based neuromodulation can eliminate catastrophic forgetting in simple neural networks. *PLoS ONE* **12**, e0187736 (2017).
85. Vecoven, N., Ernst, D., Wehenkel, A. & Drion, G. Introducing neuromodulation in deep neural networks to learn adaptive behaviours. *PLoS ONE* **15**, e0227922 (2020).
86. Ben-Iwhiwhu, E., Dick, J., Ketz, N. A., Pilly, P. K. & Soltoggio, A. Context meta-reinforcement learning via neuromodulation. *Neural Netw.* **152**, 70–79 (2022).
87. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* (MIT Press, 2018).
88. Still, S. & Precup, D. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory Biosci.* **131**, 139–148 (2012).
89. Pathak, D., Agrawal, P., Efros, A. A. & Darrell, T. Curiosity-driven exploration by self-supervised prediction. In *Proc. 34th International Conference on Machine Learning* Vol. 70 (eds Precup, D. & Teh, Y. W.) 2778–2787 (PMLR, 2017).
90. Berseth, G. et al. SMiRL: surprise minimizing reinforcement learning in unstable environments. In *9th International Conference on Learning Representations* (ICLR, 2021).
91. Burda, Y. et al. Large-scale study of curiosity-driven learning. In *7th International Conference on Learning Representations* (ICLR, 2019).
92. Carpenter, G. A. & Grossberg, S. A massively parallel architecture for a self-organizing neural pattern recognition machine. *Comput. Vis. Graph. Image Process.* **37**, 54–115 (1987).
93. Kriegeskorte, N. & Douglas, P. K. Cognitive computational neuroscience. *Nat. Neurosci.* **21**, 1148–1160 (2018).
94. de Araujo, I. E., Schatzker, M. & Small, D. M. Rethinking food reward. *Annu. Rev. Psychol.* **71**, 139–164 (2020).
95. Moerland, T. M., Broekens, J. & Jonker, C. M. Emotion in reinforcement learning agents and robots: a survey. *Mach. Learn.* **107**, 443–480 (2018).
96. Seth, A. K., Suzuki, K. & Critchley, H. D. An interoceptive predictive coding model of conscious presence. *Front. Psychol.* **2**, 395 (2012).
97. Parr, T., Pezzulo, G. & Friston, K. J. *Active Inference: The Free Energy Principle in Mind, Brain and Behavior* (MIT Press, 2022).
98. Millidge, B. Deep active inference as variational policy gradients. *J. Math. Psychol.* **96**, 102348 (2020).
99. Imohiosen, A., Watson, J. & Peters, J. Active inference or control as inference? A unifying view. In *Active Inference: IWAI 2020* (eds Verbelen, T. et al.) 12–19 (Springer, 2020).
100. Millidge, B., Tschantz, A., Seth, A. K. & Buckley, C. L. On the relationship between active inference and control as inference. In *Active Inference: IWAI 2020* (eds Verbelen, T. et al.) 3–11 (Springer, 2020).
101. Friston, K. & Ao, P. Free energy, value and attractors. *Comput. Math. Methods Med.* **2012**, 937860 (2012).
102. Sajid, N., Ball, P. J., Parr, T. & Friston, K. J. Active inference: demystified and compared. *Neural Comput.* **33**, 674–712 (2021).
103. Allen, M., Levy, A., Parr, T. & Friston, K. J. In the Body's Eye: the computational anatomy of interoceptive inference. *PLoS Comput. Biol.* **18**, e1010490 (2022).
104. Hesp, C. et al. Deeply felt affect: the emergence of valence in deep active inference. *Neural Comput.* **33**, 398–446 (2021).
105. Allen, M. & Tsakiris, M. The body as first prior: interoceptive predictive processing and the primacy of self-models. In *The Interoceptive Mind: From Homeostasis to Awareness* (eds Tsakiris, M. & De Preester, H.) 27–45 (Oxford Univ. Press, 2018).
106. Christov-Moore, L. et al. Preventing antisocial robots: a pathway to artificial empathy. *Sci. Robot.* **8**, eabq3658 (2023).
107. Zaki, J. & Ochsner, K. N. The neuroscience of empathy: progress, pitfalls and promise. *Nat. Neurosci.* **15**, 675–680 (2012).
108. Ashar, Y. K., Andrews-Hanna, J. R., Dimidjian, S. & Wager, T. D. Empathic care and distress: predictive brain markers and dissociable brain systems. *Neuron* **94**, 1263–1273.e4 (2017).
109. Feldman, R. The adaptive human parental brain: implications for children's social development. *Trends Neurosci.* **38**, 387–399 (2015).
110. Preston, S. & de Waal, F. D. Empathy: its ultimate and proximate bases. *Behav. Brain Sci.* **25**, 1–20 (2002).
111. Yoshida, N. & Man, K. Homeostatic coupling for prosocial behavior. In *ALIFE 2025: Ciphers of Life: Proc. Artificial Life Conference* Vol. 37, 32 (MIT Press, 2025).
112. Bengio, Y. & Elmoznino, E. Illusions of AI consciousness. *Science* **389**, 1090–1091 (2025).
113. Butlin, P. et al. Identifying indicators of consciousness in AI systems. *Trends Cogn. Sci.* **30**, 488–501 https://doi.org/10.1016/j.tics.2025.10.011 (2026).
114. Bernard, C. *Lectures on the Phenomena Common to Animals and Plants* (Charles C. Thomas, 1974).
115. Singh, S., Lewis, R. L. & Barto, A. G. Where do rewards come from? In *Proc. 31st Annual Conference of the Cognitive Science Society* (eds Taatgen, N. & van Rijn, H.) 2601–2606 (Cognitive Science Society, 2009).
116. Saper, C. B., Chou, T. C. & Elmquist, J. K. The need to feed: homeostatic and hedonic control of eating. *Neuron* **36**, 199–211 (2002).
117. Barto, A. G. Intrinsic motivation and reinforcement learning. In *Intrinsically Motivated Learning in Natural and Artificial Systems* (eds Baldassarre, G. & Mirolli, M.) 17–47 (Springer, 2013).
118. Levin, M. The computational boundary of a "self": developmental bioelectricity drives multicellularity and scale-free cognition. *Front. Psychol.* **10**, https://doi.org/10.3389/fpsyg.2019.02688 (2019).
119. Kaelbling, L. P., Littman, M. L. & Cassandra, A. R. Planning and acting in partially observable stochastic domains. *Artif. Intell.* **101**, 99–134 (1998).
120. Ong, S. C. W., Png, S. W., Hsu, D. & Lee, W. S. Planning under uncertainty for robotic tasks with mixed observability. *Int. J. Robot. Res.* **29**, 1053–1068 (2010).
121. Padakandla, S., Prabuchandran, K. J. & Bhatnagar, S. Reinforcement learning algorithm for non-stationary environments. *Appl. Intell.* **50**, 3590–3606 (2020).
122. Choi, S. P. M., Yeung, D.-Y. & Zhang, N. L. Hidden-mode Markov decision processes for nonstationary sequential decision making. In *Sequence Learning: Paradigms, Algorithms and Applications; Lecture Notes in Computer Science* (eds Sun, R. & Giles, C. L.) Vol. 1828, 264–287 (Springer, 2001).
123. Doshi-Velez, F. & Konidaris, G. Hidden parameter Markov decision processes: a semiparametric regression approach for discovering latent task parametrizations. *IJCAI* **2016**, 1432–1440 (2016).
124. Friston, K. et al. Path integrals, particular kinds and strange things. *Phys. Life Rev.* **47**, 35–62 (2023).
125. Barnett, L. & Seth, A. K. Dynamical independence: discovering emergent macroscopic processes in complex dynamical systems. *Phys. Rev. E.* **108**, 014304 (2023).
126. Marshall, W., Kim, H., Walker, S. I., Tononi, G. & Albantakis, L. How causal analysis can reveal autonomy in models of biological systems. *Phil. Trans. R. Soc. A.* **375**, 20160358 https://doi.org/10.1098/rsta.2016.0358 (2017).

127. Fang, X., Kruse, K., Lu, T. & Wang, J. Nonequilibrium physics in biology. *Rev. Mod. Phys.* **91**, 045004 (2019).
128. Kim, H., Valentini, G., Hanson, J. & Walker, S. I. Informational architecture across non-living and living collectives. *Theory Biosci.* **140**, 325–341 (2021).
129. Davies, P. C. W. & Walker, S. I. The hidden simplicity of biology. *Rep. Prog. Phys.* **79**, 102601 (2016).
130. Walker, S. I., Davies, P. C. W. & Ellis, G. F. R. *From Matter to Life: Information and Causality* (Cambridge Univ. Press, 2017).

Acknowledgements

This work was supported by grants IBS-R015-D1 and IBS-R015-D2 from the Institute for Basic Science (to C.-W.W. and S.J.H.), a National Research Foundation of Korea (NRF) grant (RS-2024-00398768 to S.J.H.), an Institute of Information & communications Technology Planning & Evaluation (IITP) grant (RS-2026-25518389 to S.J.H., Development of Human Cognition-Based AI Core Technology), funded by the Korea government (MSIT), and by grant HI19C1328, which is a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare (to S.L.). K.F. is supported by funding from the Wellcome Trust (226793/Z/22/Z).

Author contributions

S.L. contributed to conceptualization, investigations and visualization, and wrote the original draft. Y.O. contributed to conceptualization and investigations, and reviewed and edited the paper. H.A. and H. Y. contributed to investigations. K.J.F. carried out validation, and reviewed and edited the paper. S.J.H. contributed to conceptualization, supervision and funding acquisition, and reviewed and edited the paper. C.-W.W. contributed to conceptualization, supervision, visualization and funding acquisition, and reviewed and edited the paper.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to Seok Jun Hong or Choong-Wan Woo.

Peer review information *Nature Machine Intelligence* thanks Ismael Freire, Jeffrey Krichmar and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]Center for Neuroscience Imaging Research, Institute for Basic Science, Suwon, South Korea. [2]Department of Biomedical Engineering, Sungkyunkwan University, Suwon, South Korea. [3]Department of Intelligent Precision Healthcare Convergence, Sungkyunkwan University, Suwon, South Korea. [4]Life-inspired Neural Network for Prediction and Optimization Research Group, Suwon, South Korea. [5]Neurocore Co. Ltd, Seoul, South Korea. [6]Department of Imaging Neuroscience, Queen Square Institute of Neurology, University College London, London, United Kingdom. [7]Department of Brain Science and Engineering, Sungkyunkwan University, Suwon, South Korea. [8]Center for the Developing Brain, Child Mind Institute, New York, NY, USA. [9]These authors contributed equally: Sungwoo Lee, Younghyun Oh. ✉e-mail: hongseokjun@skku.edu; waniwoo@skku.edu