

https://dspace.org/

# Investigating the integration of interoception in a reinforcement learning pipeline

**Gebreselassie, Marrian Haileselassie**

**2025-09**
University of the Witwatersrand, Johannesburg
https://hdl.handle.net/10539/50270
Gebreselassie, Marrian Haileselassie. (2025). Investigating the integration of interoception in a reinforcement learning pipeline. [Master's dissertation, University of the Witwatersrand, Johannesburg]. WIReDSpace. https://hdl.handle.net/10539/50270

Downloaded from a DSpace repository. https://dspace.org/

# Investigating the integration of interoception in a reinforcement learning pipeline

**School of Computer Science & Applied Mathematics**
**University of the Witwatersrand**

**Marrian H. Gebreselassie**
**2633603**

**Supervised by Dr. Devon Jarvis**

**Co-supervised by Dr. Victoria Williams**

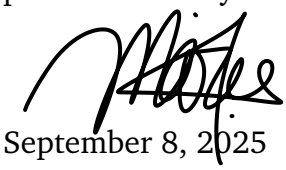
**September 8, 2025**



A dissertation submitted to the Faculty of Science, University of Witwatersrand, Johannesburg, in fulfillment of the requirements for the degree of Masters of Science

**Declaration**

I, Marrian Gebreselassie, hereby declare the contents of this dissertation in its entirety are the sole product of my efforts. This manuscript herewith is submitted in pursuit of the Masters of Science in Data Science degree from the University of Witwatersrand, Johannesburg. As such, I affirm that this work has not been submitted, proposed or presented to any other university or in supplement of any other degree.



September 8, 2025

**Abstract**

Reinforcement learning (RL), inspired by neuroscience, trains artificial agents to maximise external rewards. However, decision-making in humans involves complex evaluation and planning processes that initiate actions beyond external reward including the incentive to maintain internal homeostasis. Interoception is the brain's processing of signals from the body which convey hunger, thirst and body temperature. The initiation of actions beyond external reward is demonstrated in behaviours which address interoceptive needs like hunger. The prefrontal cortex (PFC), in coordination with other brain regions, balances responding to these intrinsically rewarding internal needs while engaging in externally rewarding activities. This study models this holistic evaluation process conducted by the PFC to manage an organism's hunger levels alongside extrinsically rewarding activities. The study draws on an analogy of a mouse in a flat room with access to a running wheel (extrinsic reward) and a block of cheese (intrinsic reward), to an RL agent in a MiniGrid environment with a green goal square and a food location. It does this by equipping an RL agent with ghrelin levels and the capacity to manage its subsequent hunger thresholds by visiting the predefined food location. The behaviour of the RL agent is consistent with the naturally observed behaviour of a hungry mouse, in that a "hungry" RL agent (i.e. when its ghrelin levels equal the hunger threshold) prioritises visiting the food location before the green goal square, emulating survival-based behaviours. By varying key parameters, an RL agent which incorporates interoception can reproduce behaviours consistent with organisms suffering from obesity and anhedonia. Hence, this study demonstrates the potential of integrating the holistic role of the PFC in human decision making onto the RL pipeline.

# Contents

**Preface**
Declaration . . . . . i
Abstract . . . . . ii
Table of Contents . . . . . iii

**1 Introduction** **1**

**2 Background and Related Work** **6**
2.1 Contextual Markov Decision Process (CMDP) . . . . . 6
2.2 Deep Q-learning . . . . . 7
2.3 Intrinsic motivation . . . . . 8
2.4 Feeding and Motivation . . . . . 11
2.5 Neuroscientific model of Food Intake as Motivated Behaviour . . . . . 13
2.5.1 Generation of motivation for food seeking and consummatory behaviour . . . . . 14
2.5.2 Maintenance of motivation for food seeking and consummatory behaviour . . . . . 15
2.5.3 Regulation of motivation for food seeking and consummatory behaviour . . . . . 17
2.6 Gaps in literature . . . . . 18

**3 Methodology** **20**
3.1 MiniGrid Environment . . . . . 20
3.2 Modelling Interoception . . . . . 21
3.3 Integrating interoceptive intrinsic reward into DQN . . . . . 26
3.4 Architecture and Hyperparameters . . . . . 27
3.5 Summary of Experiment Set-ups . . . . . 28
3.5.1 Baseline Experiments with Interoception Omitted . . . . . 28
3.5.2 Experiments with Interoception fully incorporated . . . . . 29

**4 Results** **33**
4.1 Baseline experiments with interoception omitted . . . . . 33
4.1.1 Standard RL agent . . . . . 33
4.1.2 RL agent with interoceptive signals but no food reward . . . . . 33
4.2 Experiments with Interoception fully incorporated . . . . . 35
4.2.1 Experiments investigating the influence of heightened food reward 35

4.2.2 Experiments investigating the influence of food location . . . . . 40
4.2.3 Experiments investigating the influence of hunger thresholds . . 43
4.3 Additional experiments . . . . . 46
4.3.1 Proportional Food reward . . . . . 46
4.3.2 RL agent with Ghrelin Control Actions . . . . . 49
4.3.3 Starting episode with RL agent "hungry" . . . . . 51
4.3.4 Removing effect of food . . . . . 52
4.3.5 Extreme ghrelin levels . . . . . 52
4.3.6 Arbitration of intrinsic versus extrinsic reward . . . . . 53
4.3.7 Removing food reward . . . . . 59
4.3.8 Increased function of food location . . . . . 59

**5 Discussion** **61**
5.1 Limitations . . . . . 63
5.2 Future work . . . . . 64
5.3 Concluding remarks . . . . . 65

**A Appendix 1: Hyperparameters** **66**

**B Appendix 2: Complete circuits by Averbeck and Murray [2020]** **67**

**C Appendix 3: List of Acronyms** **68**

**References** **73**

# Chapter 1

# Introduction

Interoception is the brain's processing of internal signals received from the body. These signals communicate information related to body temperature, thirst, hunger, energy levels and heart rate [Tsakiris and Critchley 2016]. This information is critical to maintain homeostasis in the body. Homeostasis is the body's "set point" of internal balance [Liu and Kanoski 2018]. The process of maintaining homeostasis on some occasions like body temperature is automatic. For example, the body releases sweat to reduce body temperature without conscious external actions by the individual. For other homeostatic signals like hunger or thirst, the individual needs to take external actions to address these needs. This study seeks to understand and model the underlying cognitive functions involved in addressing hunger to maintain homeostasis while optimising external reward.

The prefrontal cortex (PFC), in conjunction with other brain regions responsible for cognition, memory and reward processing, determines which actions an organism performs [Mohapatra and Wagner 2023]. Closely inspecting the process of decision-making in humans highlights that a multitude of factors, value calculations and intricate systems are taken into consideration before initiating actions beyond external reward [Kim 2013]. An example of this is the response to interoceptive needs. Interoceptive signals encode basic living organism needs and are the first step in initiating need-responsive actions such as eating, drinking or sleeping based on signals received from organs outside the brain [Tsakiris and Critchley 2016]. Thus, interoception can be classified as intrinsic motivation, in that the actions it motivates are not guided by external reward despite being rewarding intrinsically.

There is a clear overlap of intrinsic motivation and reward involved in homeostasis to control food intake. Although food is naturally rewarding, the initiation of food-seeking and consummatory behaviour is not solely driven by this external value but rather internal cues of hunger [Liu and Kanoski 2018; Ferrario *et al.* 2016]. In practice, a healthy organism does not solely eat because food is rewarding but because it needs to maintain energy homeostasis [Liu and Kanoski 2018]; moreover, an organism is able to moderate its eating patterns without entirely neglecting other rewarding actions. This is what enables organisms to remain satiated while fulfilling rewarding activities.

The regulation of food intake, food-seeking and consummatory behaviour has gained the attention of researchers for its interplay of dopamine (DA) as responsive to food's naturally rewarding value and its heightened value during hunger [Baik 2021; Ferrario *et al.* 2016]. Though much of the exact neural circuitry remains unclear, food intake is moderated by the brain's decision-making centre, the PFC, and demands a closer inspection of its translation into actions.

Reinforcement learning (RL) translates principles of reinforcement theory to computational agents [Sutton *et al.* 1998]. Reinforcement theory states that organisms engage in activities or actions which yield a positive reward (e.g., actions which cause an upstream release of dopamine in the brain like displays of gratitude or financial gain) [Gordan and Krishanan 2014]. The opposite is also true, in that undesirable actions are avoided because of their negative reward or punishment. This theory is applicable to the decision-making process of living organisms, wherein reward is moderated by dopamine (DA) release in the brain [Kakade and Dayan 2002] and organisms continuously engage in activities deemed pleasurable [Kakade and Dayan 2002; Gordan and Krishanan 2014]. RL adopts this reward-based framework to train artificial agents to solve tasks through positive and negative reward. The RL agent receives a positive reward for actions that solve a given task and negative rewards for actions which do not contribute to solving a given task.

This foundational approach to RL which aims to maximise external rewards has resulted in high performing agents advancing robotics and other automation applications [Nian *et al.* 2020]. In environments without immediate positive or negative rewards as feedback, RL agents struggle to solve complex tasks [Barto 2012]. For this reason, further parallels between living organisms and RL agents, beyond DA release and rewards, have led to improved algorithms for RL agents [Barto 2012; Aubret *et al.* 2019]. Barto *et al.* [2004] successfully proposed intrinsic motivation by offering RL agents intermediary intrinsic rewards unrelated to solving the environment's task and provided evidence of its added benefit to an RL agent's performance in learning reusable skills. Human curiosity is a key component of motivation which drives actions rather than external rewards [Chentanez *et al.* 2004]. By enabling agents to evaluate actions beyond external task-based reward and consider curiosity-based intrinsic reward, RL agents were able to take actions out of "curiosity" valuing characteristics like novelty, surprise, incongruity and complexity. Their incorporation of intrinsic motivation enabled RL agents to gain hierarchical skills by promoting broader environment exploration, which not only facilitated their ability to solve the given task but also to re-use these learned skills to solve different tasks not included in the agent's training [Barto *et al.* 2004].

Building on Barto *et al.* [2004]'s definition of intrinsic motivation as actions taken by an RL agent for "its own sake", interoceptive intrinsic motivation has the capacity to initiate actions which maintain the agent's homeostatic equilibrium by integrating interoceptive-based intrinsic rewards. Thus, the study seeks to answer the following questions: (1) how can interoception be incorporated into the RL pipeline?; (2) does the incorporation of interoception influence the learned policy of the RL agent compared to its complete omission?; (3) how does the modeled agent's behaviour compare

to that of a living mouse?

This study aims to replicate the regulation of intrinsically motivated interoceptive needs alongside externally rewarding tasks by modelling the holistic decision-making role of the PFC involved in the control of food intake onto an RL pipeline. The study recreates classical rodent studies on feeding behaviour by creating the analogy of a mouse in a flat room with a running wheel and block of cheese to an artificial agent in a standard MiniGrid environment containing a green goal square (extrinsically rewarding) and designated "food location" (intrinsically rewarding). The agent receives interoceptive information relaying hunger as a linear function representing its ghrelin levels. This contextual information is passed to the agent through the classical concatenation of context to state information. The RL agent's behaviour is observed and compared to the complete omission of this interoceptive information (ghrelin levels). Specifically, this study aims to understand the arbitration of intrinsic motivation versus extrinsic motivation in determining an agent's learned policy by recreating the behaviour of a hungry mouse in a flat room that contains both extrinsically and intrinsically rewarding sources (i.e., a running wheel and block of cheese). This set up ensures that both extrinsic and intrinsic reward are accessible and attainable by the artificial agent. As such, this aims to equip an RL agent, analogous to a mouse, with the capacity to engage in externally rewarding activities without entirely neglecting its interoceptive needs.

Using a series of structured control experiments, the study deploys an artificial RL agent with the ability to manage its hunger levels by visiting a specific "food location" within the grid representing food that lowers its ghrelin levels (i.e. analogous to the mouse eating the block of cheese). When the agent's ghrelin levels exceed a pre-defined hunger threshold (making the agent "hungry"), the agent leaves the externally rewarding green goal square to "eat" at the food location, by taking a few steps at the specified food location on the grid. These results are further compared to an RL agent equipped with actions that autonomously control its ghrelin levels without needing to visit the "food location". With varying key parameters, an interoception-sensing RL agent can exhibit obesogenic behaviours of excessive "eating" unaccompanied by hunger and anhedonia-like behaviours which neglect internal hunger cues entirely [Zigman *et al.* 2016; Murray *et al.* 2021].

The main contributions of the research are:

1. Incorporating interoception in RL: Interoceptive information relaying hunger by ghrelin levels was successfully modeled in an RL agent. It was incorporated as contextual information which influences the RL agent's learned policy formed by a Contextual Markov Decision Process (CMDP). Ghrelin levels affect the reward functions associated with state transitions in the environment based on whether the agent is below or above the set hunger threshold, evoking baseline and heightened food reward values respectively.

2. Interoceptive-driven policy formation: Through the incorporation of interoceptive information, an artificial RL agent is able to learn a policy which attains external reward and manages its interoceptive needs. This is demonstrated by an agent

consistently learning a policy which intersects the food location enroute to the goal square (i.e., while satiated) tested with different food locations on the grid. During a single episode, this agent leaves the goal square to visit the pre-defined food location when its hunger threshold is exceeded and it receives a heightened food reward.

3. Biologically relevant behaviour: RL agents which incorporate interoceptive information behave like living organisms in prioritising interoceptive needs when these agents are deployed at a deficit. This is demonstrated by the RL agent first visiting food locations within the grid before the green goal square when deployed at the threshold hunger level tested with different food locations. Similar behaviour is observed in RL agents with autonomous ghrelin control actions when deployed at deficit, in that these agents first lower their ghrelin levels below the set hunger threshold before proceeding to the green goal square. The results obtained demonstrate the role of toggling between baseline and heightened food reward in recreating healthy feeding patterns observed in living organisms onto RL agents.

4. Modelling pathological behaviours: The model reproduces dysregulated food-reward processing analogous to obesity and anhedonia with RL agents which possess interoception when the pre-defined food location is distant from the goal square, hunger thresholds are extremely low, and heightened food reward value is comparable to the base environment reward respectively. This is demonstrated by RL agents who learn a policy which terminates at the food location, and never venture to the far-away goal square even when their ghrelin levels are below the hunger threshold (i.e. exhibiting obesity-like behaviour). This is also demonstrated by RL agents who remain at the green goal square despite exceeding the hunger threshold (i.e. anhedonia-like behaviour).

Although prior work has successfully incorporated intrinsic motivation, it does so to improve an agent's exploration of the environment and supplement the development of skills to better navigate this environment ultimately to maximise extrinsic task-based reward [Aubret *et al.* 2019]. These definitions of intrinsic motivation overlook the homeostasis and self-maintenance of the RL agent. Thus, a key contribution of this work is the expanded definition of intrinsic motivation in RL agents through the integration of interoception, which enables agents to balance their internal needs while attaining external rewards. Overall, the study contributes to analogies between humans and RL agents, moving away from extrinsic reward-based frameworks to more holistic neuroscientific-based models.

The following is an overview of subsequent chapters: Chapter 2 details the literature and theoretical concepts of intrinsic motivation in reinforcement learning and the state-of-the-art understanding of neural circuitry underlying food intake which ground this work; Chapter 3 provides details of the key parameters utilised to model interoception and the different experimental set ups to investigate the influence of each parameter; Chapter 4 provides the obtained results from the collection of conducted experiments including commentary on emerging patterns of behaviour observed and alternative experiments with autonomous ghrelin control actions; and finally Chapter 5 provides a

detailed discussion on the limitations of the results obtained, a thorough commentary on the findings of this study, potential avenues for future work as well as concluding remarks on the contribution of this work.

# Chapter 2

# Background and Related Work

This study investigates the integration of interoception, the brain's sensation of one's internal organs observed in living organisms, onto an RL pipeline. In particular, it models hunger and consequent food seeking and consummatory behaviour [Levine and Billington 1997]. These behaviours have been studied extensively in the rodents (mice) [Jensen *et al.* 2013; Heinrichs 2001]. The study builds on three main avenues: (1) intrinsic motivation in artificial agents trained with reinforcement learning, (2) rodent studies investigating feeding behaviours, and (3) decision-making pathways regulating food intake in living organisms.

## 2.1 Contextual Markov Decision Process (CMDP)

RL environments assume a Markov Decision Process (MDP) where an agent performs actions in an environment which elicits states with the aim of maximising rewards. An MDP is defined by a tuple $\langle S, A, T, R, \gamma \rangle$ where:

- S - the set of possible states;
- A - set of possible actions which can be taken by the agent;
- P - the transition function P : S×A×S $\rightarrow$ P($S'$ |S, A) – the next state evoked by an agent performing an action in a given state
- R : S × S × A $\rightarrow \mathbb{R}$ - the reward function is the outcome obtained by agent after performing action A at a given state, S
- $\gamma \in$ [0, 1] - the discount factor is a hyper-parameter to balance between long and short-term rewards when an agent selects which actions to take
- $\rho_0$ : S $\rightarrow$ P(S) the initial distribution of states

Within this framework, the goal of an RL agent is to develop a policy, $\pi$, which associates actions to states to obtain an optimal policy $\pi^*$ which maximises reward [Sutton *et al.* 1998]. After forming an optimal policy, RL agents struggle to accommodate changes in

the environment or state dynamics [Aubret *et al.* 2019]. This has led researchers to investigate the addition of contextual information in the formation of an agent's optimal policy.

Previous studies have shown integrating contextual information through concatenation yields positive results for RL agents' performance [Lee *et al.* 2020; Ball *et al.* 2021]. In this setting, the RL agent assumes a Contextual Markov Decision Process (CMDP) which considers differences in context when mapping actions to states. A CMDP is a tuple, $\langle C, S, A, M', \gamma \rangle$, where C, S and A are the context, state and action spaces respectively. Each context $c \in C$ is mapped to an MDP by the following function, $M' = \langle S, A, T_c, R_c, \gamma \rangle$ . Thus, each context results in unique reward ($R_c$) and transition ($T_c$) functions [Hallak *et al.* 2015]. This study assumes a CMDP where different ghrelin levels represent different contexts of the agent. In experiments with heightened food reward described in Chapter 3, the ghrelin level of the agent (i.e., the context) determines the reward function associated with state transitions in the environment.

## 2.2 Deep Q-learning

Following a CMDP, the agent requires a learning function which estimates the value of state-action transitions. This study utilises deep q-learning. As defined by Mnih *et al.* [2013], a deep q-network (DQN) RL agent seeks to maximise future rewards. The base assumption is that future rewards are discounted by some factor, $\gamma$ (gamma) at each step. A DQN agent learns to maximise discounted, cumulative reward, $R_{t0}$, defined as:

$$R_{t_0} = \sum_{t=t_0}^{\infty} \gamma^{t-t_0} r_t$$

- where $r_t$ is the reward received at time step (t) by the agent for executing some action (a) in state (s)

Every state has a value (V(s)) which is the expected reward, r, from state, s, under policy $\pi$. For every action, a, taken in state, s, there is an expected reward denoted as the action-value function, Q(s,a). Thus, the optimal action-value function Q*(s,a) represents actions which maximise future rewards and by extension form an optimal policy, $\pi^*$. The optimal policy is defined as:

$$\pi^*(s) = \arg\max Q^*(s, a)$$

- where $\pi^*$ is a policy that maximises rewards

Q*(s,a) obeys the Bellman equation. The Bellman equation evaluates a state depending on the reward achieved for that given state, s, and the value (total possible rewards) of the next state, $s'$. This is defined as:

$$Q^{\pi}(s, a) = r + \gamma Q^{\pi}(s', \pi(s'))$$

The temporal difference error is used to form more accurate estimates of Q(s,a) for

every possible state-action pair in the environment. It does this by computing the difference between predicted and actual reward at each state, s, for a given action, a. $\delta$, is defined as:

$$\delta = (r + \gamma \max Q(s', a)) - Q(s, a)$$

DQN possess a replay memory which stores a collection of the agent's interactions with the environment at each step. This stores the state, action taken, next state and reward obtained by the RL agent for a fixed number of transitions. These transitions are pooled over many episodes into a replay memory. These transitions are used to calculate the loss and update q-value estimates by comparing against actual reward obtained for state-action transitions. A random sample from the replay memory forming a batch, B, is used to calculate the loss of the network and optimise its node weights.

The temporal difference error is minimised using the mean squared error (MSE), defined as:

$$\mathcal{L} = \frac{1}{|B|} \sum_{(s,a,s',r) \in B} (\delta^2)$$

where:

- B – batch of sampled transitions
- $|B|$ - batch size
- $\delta$ – temporal difference error

Deep Q-learning was selected for its model-free (purely trial-and-error learning) nature whereby an agent learns about the environment during training [Sutton *et al.* 1998; Mnih *et al.* 2013]. This allows for flexibility in incorporating additional characteristics such as food location and food reward in the environment detailed in Chapter 3. Secondly, the DQN allows for off-policy optimisation which allows an agent to take actions part of a sub-optimal policy. Thus, an agent can take actions which are not part of the optimal policy and this is particularly helpful in enabling the agent to adequately explore the environment in search of the food location detailed in Chapter 3.

## 2.3 Intrinsic motivation

Intrinsic motivation was introduced to RL applications to support agent learning in reward scarce environments through a number of approaches [Barto *et al.* 2004; Kayal *et al.* 2025; Aubret *et al.* 2019]. Latyshev and Panov [2024] synthesise intrinsic motivation as an extension of analogies between RL agents and the psychology of motivation in human behaviour; hence, the generation of a reward signal dependent on "inherent satisfaction" rather than an extrinsic task. They propose a framework summarising existing intrinsic motivation applications as: (1) complementary intrinsic reward; (2) exploration policy, and (3) intrinsically motivated goals [Latyshev and Panov 2024].

These applications utilise an intrinsic reward signal generated from the agent's model of the environment to promote exploration and the development of task-agnostic skills towards maximising external rewards. This study aligns with intrinsic motivation approaches that utilise a complementary intrinsic reward defined as:

$$r_{\text{total}} = r_{\text{ext}} + \alpha * r_{\text{int}}$$

where:

- $r_{\text{ext}}$ – extrinsic reward
- $\alpha$ – adaptive coefficient
- $r_{\text{in}}$ – intrinsic reward

Kayal *et al.* [2025] reach similar conclusions about current applications of intrinsic motivation as either "knowledge-based" or "competence-based" and further explore the different kinds of intrinsic reward signals to support agent learning. Their work notes that existing applications of intrinsic motivation encourage the agent to attain policy, state or behavioural diversity to improve performance yet many of these applications fail to investigate the impact of different kinds of diversity on the agent's behaviour [Kayal *et al.* 2025]. Their work systematically examines four intrinsic rewards: (1) State Count, (2) Intrinsic Curiosity Module (ICM), (3) Maximum Entropy, and (4) Diversity is All You Need (DIAYN) which each correspond to a form of diversity with different MiniGrid environments. Their experiments assume a complementary intrinsic reward signal similarly defined as:

$$r_{\text{total}} = r_{\text{ext}} + \beta * r_{\text{int}}$$

where:

- $r_{\text{ext}}$ – extrinsic reward
- $\beta$ – intrinsic reward coefficient
- $r_{\text{in}}$ – intrinsic reward

Thus, existing applications of intrinsic motivation offer proxies of promoting exploration, by generating intrinsic reward signals that reward the agent for entering unexpected states, information gain or gaining unsupervised skills which may not be useful in solving the environment's task [Kayal *et al.* 2025; Latyshev and Panov 2024]. Although this work utilises a complementary intrinsic reward signal, the formation of the intrinsic reward is not geared towards promoting exploration of the environment and maximising external reward. Instead, this study proposes an intrinsic reward signal that reflects the "inherent satification" or dissatisfaction associated with eating and hunger respectively.

Barto *et al.* [2004] was among the first to argue that if RL agents are meant to imitate living organisms, then external rewards as the single motivation of actions emulated in the RL environment is an inaccurate depiction. Instead, their work presented internal and external environments which are considered by an organism when making decisions depicted in Figure 2.1. Their work explored intrinsic motivation such as curiosity

in humans which valued novelty, surprise, incongruity, and complexity as positively rewarding in RL-agents compensated by intrinsic reward [Barto *et al.* 2004]. Their work enabled the collection of hierarchical skills which resulted in higher learning speed.

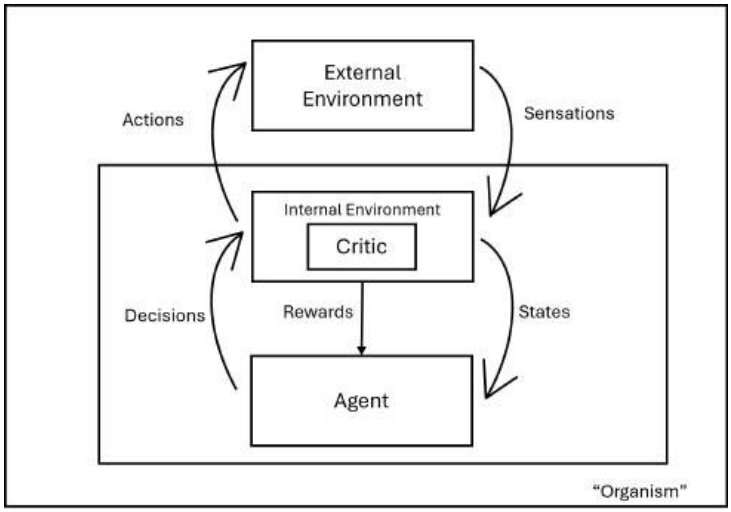



Figure 2.1: Factoring of internal and external environment which influence decision making in living organisms as explained by Barto *et al.* [2004]

Other than developing broad re-usable skills, other research into intrinsic motivation has explored equipping agents with the autonomy to select new goals and tasks to enable their generalisation to different environments as mentioned by Kayal *et al.* [2025]. Merrick [2012] proposed utilising intrinsic motivation and introspection to continuously update the goals and tasks which RL agents pursue. Their work builds on notions of curiosity and internal reward presented by Barto *et al.* [2004] and present an alternative approach to generalisability. However, a key distinction between interoception and introspection must be established. Introspection defines reasoning about one's feelings, thoughts and mental state. When applied to RL agents, this enables the agent to pursue policies which favour exploration rather than extrinsic reward. On the other hand, interoception refers to signals sent from internal organs to the brain; an example is feeling thirsty. Although these organs are external to the brain, they influence the consequent actions initiated by the brain and in turn, the brain experiences a reward (i.e., dopamine release) for responding to the signals communicated from the relevant organ system [Baik 2021].

Aubret *et al.* [2019] present a comprehensive mapping of the uses of intrinsic motivation in reinforcement learning. Their work distils two key areas wherein intrinsic motivation is employed: (1) knowledge acquisition and (2) skill learning [Aubret *et al.* 2019]. Both avenues rely on an agent's exploration of the environment to either gain a comprehensive understanding of its external environment (which relies on accurate state representations) and collection of reusable abstract skills for internal reward (which contribute to quicker attainment of external reward) respectively [Aubret *et al.* 2019]. These broad sets of intrinsic motivation do not reflect intrinsic reward gained for actions which address the agent's internal needs similar to a living organism's physiological interoceptive needs like eating, drinking or sleeping.

## 2.4 Feeding and Motivation

Researchers conduct ethical experiments on mice to study neuroscientific behaviours such as social interaction, feeding and cognition. The selection of rodents is because of their physiological similarity to humans (i.e. both are mammals with similar organ systems) and the ability to readily mutate variants of mice with preset medical conditions or lacking specific neuroreceptors for experiments [Perlman 2016; Jensen *et al.* 2013; Heinrichs 2001]. Mice subjects are fasted prior to experiments to ensure that natural satiety levels during tests are unaltered from a previous meal [Jensen *et al.* 2013]. Heinrichs [2001] provides a definition of hunger as "an internal state which represents nutritional depletion and energises food consummatory behaviours". In general, fasted mice are made to conduct tasks such as pressing a lever to attain food and placed with salient objects within their flat rooms including a running wheel (which is naturally rewarding for mice) to observe their behaviour [Greenwood *et al.* 2011]. Greenwood *et al.* [2011] found voluntary wheel running activates the mesolimbic pathway, also known as the reward system of the brain, in mice. The mesolimbic pathway consists of dopamine projections from the ventral tegmental area (VTA) (in the mid brain) to other brain regions like the ventral striatum (which includes the nucleus accumbens (NAcc)), PFC, amygdala and hippocampus [Lewis *et al.* 2021; Greenwood *et al.* 2011].

Consequent studies investigating the brain circuitry and hormones involved in food intake have extensively been conducted on mice. Early works by Kelley *et al.* [2005] offered significant findings on the role of exciting and inhibiting different brain regions, specifically the NAcc, in eliciting feeding behaviours in rats. The NAcc reflects the desirability or reward of an action and plays a key role in defining motivated behaviours to advance pleasurable activities [Floresco 2015]. Although classical views have sought to deconstruct distinct homeostatic and hedonic brain circuits regulating food intake, Liu and Kanoski [2018] detail the overlapping functions, interactions and influence of hormones and neural pathways involved in feeding based on experiments conducted on mice. Their work underscores that regulating food intake and energy balance is unlike other interoceptive homeostatic functions like regulating body temperature, in that it requires conscious goal-directed consummatory behaviour initiated by the brain in response to signals received from the stomach [Liu and Kanoski 2018; Ferrario *et al.* 2016]. Although studies in mice have made significant progress in understanding food intake, the route of overlapping of homeostatic and hedonic circuits converting hunger sensation to food-seeking strategies is yet to be fully understood [Kelley *et al.* 2005; Liu and Kanoski 2018; Simpson and Balsam 2015].

Food seeking and consummatory behaviour requires motivation to guide its subsequent actions. Kim [2013] provides a three-part process of generating, maintaining and regulating motivation for different behaviours including eating shown in Figure 2.2. Their proposed model is complimentary to reinforcement theory and the theory of intrinsic motivation which states that organisms are constantly motivated to obtain rewards [Simpson and Balsam 2015]. Their work provides a detailed review of the different prefrontal cortex (PFC) regions in conjunction with other brain areas involved in the evaluation of actions relevant to motivation. In particular, Kim [2013]'s work elabo-

rates on the subsystem of the orbitofrontal cortex (OFC) (which is part of the PFC), amygdala and NAcc which conducts a comprehensive value assessment of potential actions. Kim [2013] also details that the dorsolateral prefrontal cortex (dlPFC), medial prefrontal cortex (mPFC) and anterior cingulate cortex (ACC) which are all part of the PFC are primarily involved in the formation and execution of plans to achieve goals.

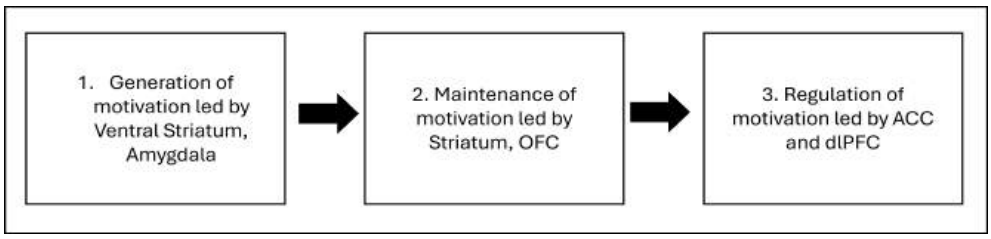



Figure 2.2: Main processes and key components of neuroscientific model of motivation proposed by Kim [2013] OFC – orbitofrontal cortex; ACC – anterior cingulate cortex; dlPFC – dorsolateral prefrontal cortex

Averbeck and Murray [2020] provided further details on the neurocircuitry underlying need-responsive behaviours in living organisms. Their work detailed the key role of the hypothalamus in controlling behaviours directly associated with survival like eating, drinking and reproduction. Their work proposes two distinct neural circuits – ventral shown in Figure 2.4 and dorsal shown in Figure 2.3 – involved in learning and goal-directed behaviour which are activated during eating. The full circuitry can be found in Annex B. According to their work, the ventral circuit is connected to the hypothalamus and defines goals of the organism relevant to basic needs. The dorsal circuit has minimal connectivity to the hypothalamus and flexibly defines an action plan to respond to these goals [Averbeck and Murray 2020].

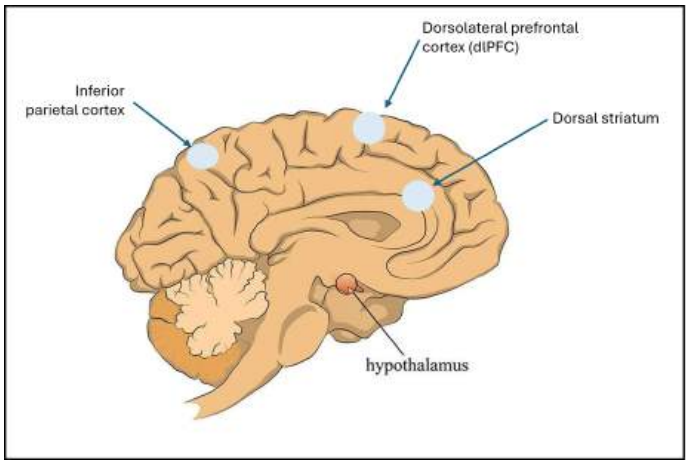



Figure 2.3: Simplified version of dorsal circuit detailed by Averbeck and Murray [2020] which develops plan on-the-go to respond to goals.

Food intake is part of an internal process to maintain energy homeostasis governed by the hypothalamus [Ferrario *et al.* 2016]. The hypothalamus functions as a link between the body's central nervous system (which controls cognitive function and movement)

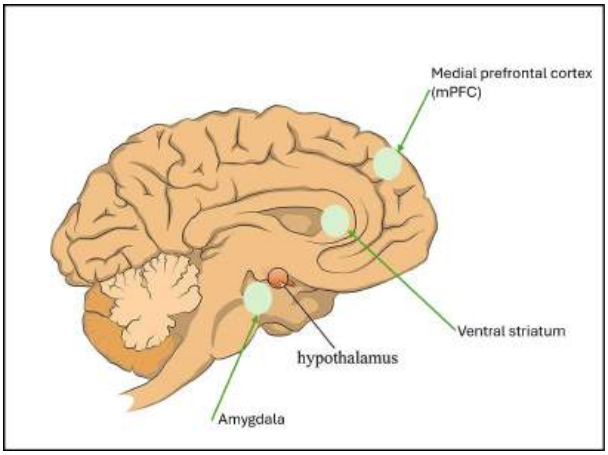



Figure 2.4: Simplified version of ventral circuit detailed by Averbeck and Murray [2020] which defines goals.

and endocrine system (which regulates hormones and bodily functions) [Saper and Lowell 2014]. Numerous researchers agree that food intake is initiated by ghrelin secretion in the stomach when it is empty [Ferrario *et al.* 2016; Skibicka *et al.* 2011]. Skibicka *et al.* [2011] injected ghrelin directly into the VTA in mice which increased motivated behaviour for sucrose reward. Their work also found that microinjection of ghrelin to the VTA and NAcc increased free feeding of regular pellet food in rats. Nieh *et al.* [2016] reached similar conclusions about the impact of ghrelin on the VTA. Their work demonstrated that when ghrelin is released in the stomach, it acts on the arcuate nucleus (ARC) in the lateral hypothalamus (LH). This causes the LH to project inhibitory GABAergic projections to the VTA, which causes the upstream release of dopamine in the NAcc [Nieh *et al.* 2016].

These findings show that food is naturally rewarding (i.e., causes a release of DA), and it also addresses a basic physiological interoceptive need of hunger (i.e., ghrelin released in the gut) that is crucial to the survival of the organism. Thus, motivation to obtain food may not always be driven by the reward associated with food itself. For this reason, neuroscientific research on food consumption intersects known concepts of hedonic, motivated behaviour associated with the NAcc and VTA, as well as neural circuits involved in homeostatic control associated with the LH and ARC [Stover *et al.* 2023]. In conclusion, multiple studies investigating food intake (eating) concur that the NAcc is activated as an upstream consequence of increased ghrelin levels during hunger and this represents the overlap of homeostatic and mesolimbic pathways [Averbeck and Murray 2020; Nieh *et al.* 2016; Ferrario *et al.* 2016; Stover *et al.* 2023].

## 2.5 Neuroscientific model of Food Intake as Motivated Behaviour

This study seeks to mirror the overlap of homeostatic and hedonic circuits in food intake onto artificial RL agents. It builds on circuitry and pathways proposed by Aver-

beck and Murray [2020] and Kim [2013]. The following sections in line with Figure 2.2, provide a detailed description of: (2.5.1) detection of hunger as the generation of motivation related to food-seeking behaviour, (2.5.2) interactions of various hormones with brain regions to maintain motivation and initiate motivated food-seeking behaviour, and (2.5.3) post-meal hormones and neural interactions to cease consummatory behaviour in a healthy organism.

### 2.5.1 Generation of motivation for food seeking and consummatory behaviour

Hunger and food consumption are consequences of underlying neural circuits which regulate calorie intake, glucose metabolism and energy expenditure in the body [Ferrario *et al.* 2016; Farr *et al.* 2016; Nieh *et al.* 2016]. This is modulated by two primary groups of hormones, neurons and neurotransmitters that either increase appetite while decreasing energy expenditure or decrease appetite while increasing energy expenditure. The following is a step-by-step process of hunger sensation in the stomach to its detection in the brain as the generation of motivation to eat.

#### Homeostatic circuit: Ghrelin signals hunger

Ghrelin is a hormone released in the stomach when it is empty, i.e. during hunger [Abdalla 2015]. Presence of ghrelin in the stomach is detected and communicated by the gastric vagal afferent pathway to the brain [Date *et al.* 2002]. This pathway terminates at the nucleus solitary tract (NTS) which delivers visceral signals to the hypothalamus [Stover *et al.* 2023]. The NTS serves as a connection between external organ systems (i.e., the stomach in this case) and the brain. Without this pathway, hunger cues would not reach the decision-making centre of the brain to initiate food-seeking behaviour. Within the hypothalamus, ghrelin secretion affects the ARC. The ARC contains two main neurons – orexigenic and anorexigenic. Orexigenic neurons promote appetite and decrease energy expenditure; these are activated by ghrelin in the hunger state. Anorexigenic neurons suppress appetite and increase energy expenditure; these are activated by leptin in the satiated state [Ferrario *et al.* 2016]. During states of hunger, leptin levels are decreased as ghrelin increases to promote food-seeking behaviour.

#### Overlap of homeostatic and hedonic circuits

As detailed in section 2.4, ghrelin increases VTA dopamine release. The LH affected by ghrelin secretion, inhibits the VTA causing the release of dopamine in the NAcc, a major part of the ventral striatum. This dopamine released by the VTA activates the mesolimbic system connected to the NAcc, septum, amygdala and the hippocampus. This activation of dopamine (DA) neurons in NAcc has been observed to increase "sucrose-seeking" behaviour and feeding behaviour in general as well as motivational salience across different settings [Ferrario *et al.* 2016; Skibicka *et al.* 2011; Farr *et al.* 2016; Nieh *et al.* 2016; Timper and Brüning 2017].

The activation of DA neurons from the VTA corresponds to the generation of motivation phase proposed by Kim [2013]'s model which is the first step of initiating motivated behaviour. The involvement of the hypothalamus, VTA and NAcc aligns with Averbeck and Murray [2020]'s ventral circuit which interprets information from the hypothalamus about ghrelin levels and defines food-seeking goals. These components align with the core sub-processes of the generation phase proposed by Kim [2013] which are reward-driven approach (pursuing actions previously associated with reward), reward anticipation (the calculation of an expected reward from current actions) and incentive salience (increased desirability and attention to stimuli associated with reward). In conclusion, the release of dopamine in the NAcc activates the mesolimbic reward system and triggers motivated food-seeking behaviour.

### 2.5.2 Maintenance of motivation for food seeking and consummatory behaviour

The activation of the NAcc, known as the pleasure centre of the brain, in the hunger state represents the overlap between homeostatic and hedonic systems guiding decision-making related to food-seeking behaviour [Liu and Kanoski 2018; Ferrario *et al.* 2016; Timper and Brüning 2017; Baik 2021]. Based on Averbeck and Murray [2020] and Kim [2013] models, this study synthesises that decisions pertaining to food consumption are moderated as motivated-behaviours governed by the mesolimbic and mesocortical pathways which evaluate actions, outcomes and rewards to make subsequent decisions. The hypothesis of this model is supported by the activation of the NAcc by the LH-VTA GABAergic projections and the established role of NAcc in other motivated-behaviours.

#### Mesolimbic and Mesocortical Pathways

The mesolimbic dopamine system is responsible for reward anticipation and learning, whereas the mesocortical dopamine system leads encoding the relative value of the reward and goal-directed behaviour [Kim 2013]. The mesocortical dopamine system connects the medial prefrontal cortex (mPFC), the anterior cingulate cortex (ACC), and the perirhinal cortex. Key sub-processes to maintain motivation are: (1) value-based decisions (continuous evaluation of action-outcomes), (2) associative learning (formation of stimulus-action-outcome patterns) and (3) positive reward-prediction error (when actual reward attained from actions is higher than predicted reward value). These processes rely on assigning and recalling the abstract value of rewards, actions and their outcomes to guide decision-making. Aside from dopamine neurons, the orbitofrontal cortex (OFC), ACC and hippocampus encode reward values and associations which inform the PFC's decision-making process of motivated behaviours [Kim 2013].

#### Evaluating actions, rewards and outcomes

Dopamine neurons encode the reward error prediction (RPE) of an action [Kim 2013; Baik 2021]. RPE is the difference between expected and actual reward outcomes. When RPE is positive, this means the outcome is better than the expected reward and vice versa. Larger positive RPE leads to improved learning which associates cues to rewards

hence promoting the reward-obtaining actions. Overtime, repeated rewards lead to lower RPE, thus declining motivation for the previously rewarding action as the brain reduces dopamine release [Kim 2013].

### The prefrontal cortex (PFC) as the final arbitrator

The prefrontal cortex (PFC) is the final decision-maker of eating behaviours, informed by its other functions related to cognition and learning [Mohapatra and Wagner 2023; Chau *et al.* 2018]. The PFC is crucial for orchestrating behaviour for both reward-driven and emotion-driven behaviours, including food consumption [Chau *et al.* 2018]. Within the prefrontal cortex (PFC) 'reward sensitive' processes occur at the dorsal anterior cingulate cortex (dACC), ventromedial prefrontal cortex (vmPFC) and orbitofrontal cortex (OFC). These areas are also connected to the striatum, which includes the NAcc of the ventral striatum. These areas are directly involved in reward-based decision-making of motivated behaviours. These regions within the PFC contribute to attributing value to impulses, associating that value with actions and deciding between different actions. This network between the PFC (e.g., OFC) and mesolimbic circuit (e.g., NAcc) is central to reward-based decision-making involving food intake [Chau *et al.* 2018].

According to Kim [2013]'s motivational process (Figure 2.2), the orbitofrontal cortex (OFC) and striatum are the key maintainers of reward through computing various markers of holistic action-outcome value. The OFC, amygdala, and NAcc are central to reward processing [Kim 2013; Rolls 2023]. The OFC conducts a go/no-go assessment based on the expected outcome of each action (i.e., determining which actions to execute or avoid respectively) and uses the output to inhibit the NAcc. When the OFC-amygdala-NAcc subsystem completes the holistic value calculation of different action-outcome pairs, the decided actions are planned for and executed by other PFC areas in conjunction with parts in the striatum [Kim 2013; Rolls 2023].

### Translating motivation to actions

Based on the evaluation of the OFC-Amygdala-NAcc subsystem, other PFC areas formulate and execute a food-seeking plan to respond to hunger signals. The dorsal circuit described by Averbeck and Murray [2020] flexibly defines an action plan to respond to food-seeking goals formed by the ventral circuit which is connected to the hypothalamus. Averbeck and Murray [2020]'s ventral and dorsal circuits interact at the PFC, where the ventral lateral PFC (vlPFC) and ACC lie. Altogther, this hints that the ACC is where decisions are converted into motor-based actions [Chau *et al.* 2018]. The dorsal striatum in the dorsal circuit (refer to Figure 2.3), projects to the substantia nigra which supports exploratory behaviour [Chau *et al.* 2018]. This is in line with findings by Nieh *et al.* [2016] that activating the GABAergic LH-VTA projection promotes investigation of the most proximal salient objects observed in rodent studies. Another component of Averbeck and Murray [2020]'s dorsal circuit is the dorsal lateral prefrontal cortex (dlPFC) which is part of the PFC and is heavily involved in reward-based decision making of motivated behaviours (refer to Figure 2.3). This aligns with Kim [2013]'s model,

where the dlPFC (i.e., by extension the PFC) is the key regulator of motivation and motivated behaviours.

### 2.5.3 Regulation of motivation for food seeking and consummatory behaviour

#### Motor execution of food-seeking behaviours

The dorsal anterior cingulate (dACC) which is part of the ACC and presupplementary motor area (pre-SMA) execute motor plans to initiate food-seeking behaviour determined by the PFC in conjunction with other brain regions [Chau *et al.* 2018]. The connectivity of the dACC positions it ideally to relay value-based decisions determined by the OFC-amygdala-NAcc subsystem depicted in Figure 2.5. The dACC is connected to nearby areas of the PFC, such as the dlPFC, and adjacent ACC areas. The dACC itself is directly connected to majority of the striatum (which includes the NAcc), as well as the amygdala. Given its rich connections, the dACC may impact, and be impacted by, dopamine released from mesolimbic structures. Its direct connections to the pre-SMA also allows it to directly influence motor output. Altogether, the dACC is located at an important intersection between the brain's reward valuation networks and their conversion to action [Chau *et al.* 2018].

The involvement of the dACC and dorsal circuit proposed by Averbeck and Murray [2020] fit accurately with the key components of Kim [2013]'s motivation regulation phase arbitrators – the anterior cingulate cortex (ACC) and the dorsolateral prefrontal cortex (dlPFC). The core sub-processes in this phase are goal-directed control, cognitive control and negative reward-prediction error.

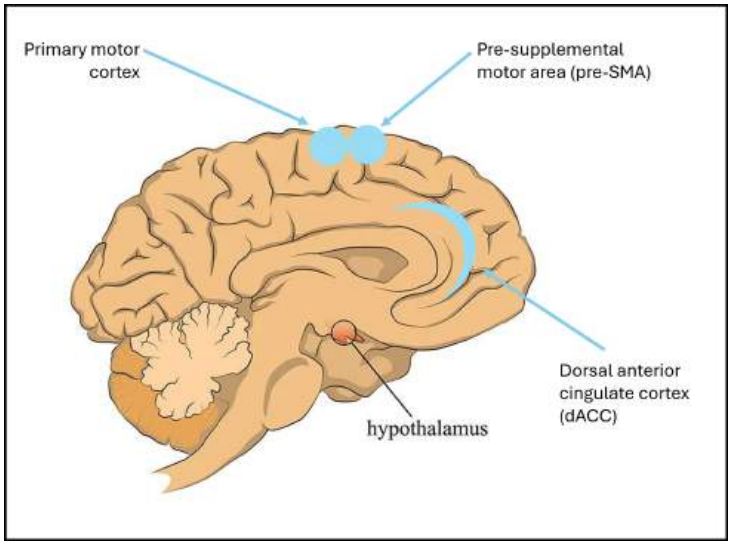



Figure 2.5: Illustration of the proximity of the dACC and pre-SMA within the PFC as potential translation point of decisions to actions.

### Shift to satiated state

After a meal, ghrelin levels decrease while leptin levels and energy expenditure increase. Paired with the repeated reward of food while eating, the RPE decreases, and dopamine release expressed by food consumption decreases until the action of feeding is no longer actively motivated and halted. In conclusion, meal termination is brought about by the shifting ghrelin and leptin hormone levels paired with decreased dopamine release by food consumption (i.e., eating is no longer as rewarding as when hungry). This moves the body from an internal state of hunger to satiety and demotivation of feeding behaviours. A compiled summary illustration of the processes described in sections 2.5.1, 2.5.2 and 2.5.3 is shown in Figure 2.6.

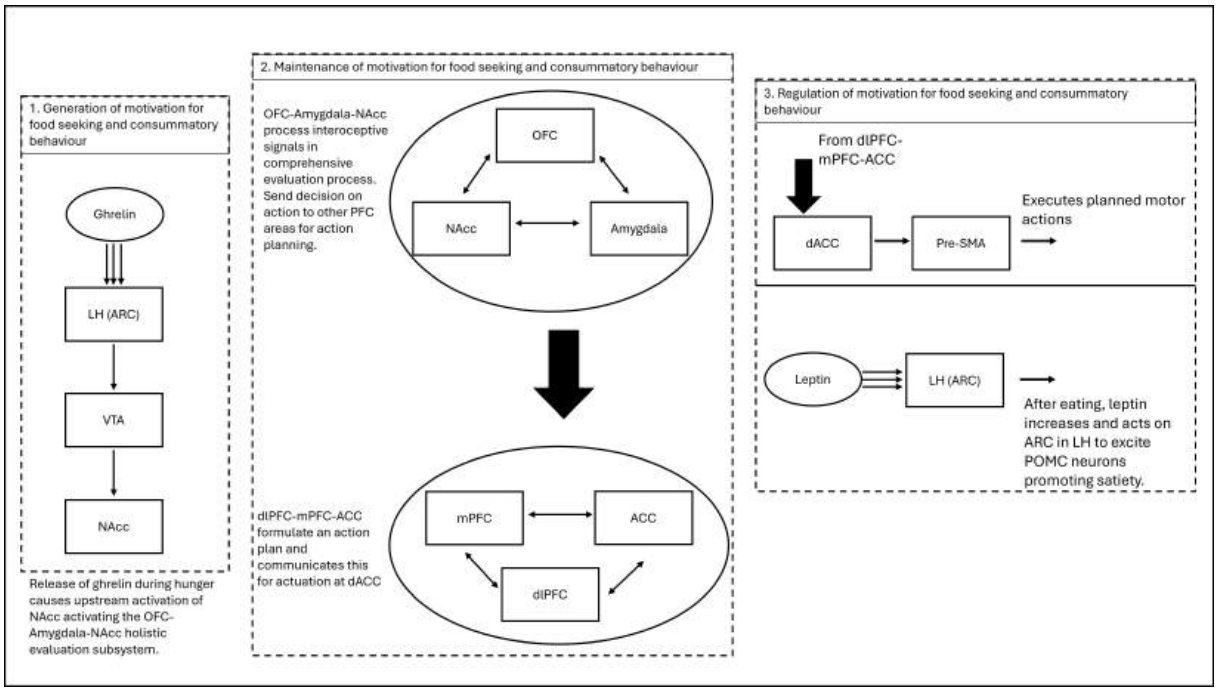



Figure 2.6: Overall pipeline of this study divided as per Kim [2013]'s model of motivation (1. Generation; 2. Maintenance; 3. Regulation) applied to food intake incorporating key brain regions described in Averbeck and Murray [2020]'s ventral and dorsal circuits involved in regulating food intake.

## 2.6 Gaps in literature

A definitive trajectory of overlapping homeostatic and hedonic circuits involved in regulating food consumption in humans is unclear due to limitations associated with accurately mapping dopaminergic trajectories within the brain, the physiological differences between humans and rodents species and the complex interconnectedness of different regions within the brain [Baik 2021; Chau *et al.* 2018]. Regardless, the role of both circuits in maintaining energy homeostasis, which is crucial to the survival of living organisms, is acknowledged [Averbeck and Murray 2020; Baik 2021; Ferrario *et al.* 2016; Liu and Kanoski 2018].

The traditional RL pipeline deploys agents which prioritise maximising extrinsic reward and utilsing intrinsic motivation to promote exploration and develop skills which contribute to this ultimate extrinsic aim in new environments [Sutton *et al.* 1998; Aubret *et al.* 2019]. RL has yet to model and reflect the interplay of homeostatic and hedonic circuits onto artificial agents despite its integration of intrinsic motivation. Thus, the relevance of this work is that it lies at the understudied intersection of food intake as a motivated behaviour and replicating the neurological principles of energy homeostasis onto RL agents. This study explores the potential of integrating an intrinsic reward signal that incentivises an RL agent to address its internal needs (modeled by ghrelin levels observed in humans) rather than exploring the environment as other applications of intrinsic motivation.

# Chapter 3

# Methodology

Based on the theoretical framework outlined in Chapter 2, this study models food intake as a motivated-behaviour (i.e., internally driven to pursue the external action of eating [Simpson and Balsam 2015]) building on circuitry described by Kim [2013] and Averbeck and Murray [2020] illustrated in Figure 2.6. The main hypothesis of this study based on the role of dopamine release in increasing food seeking behaviour during hunger, is that the holistic valuation of food seeking behaviour conducted by the OFC-Amygdala-NAcc subsystem detailed in section 2.5.2 is influenced by the organism's internal state of hunger. This means that when an organism is hungry, the brain perceives actions related to food as having a higher holistic value and are prioritised in subsequent action planning. The emphasis on "holistic value" is intentional to reflect that the reward value of food is not just a function of how naturally rewarding food is, but this value is also a function of its importance to the survival of the organism [Averbeck and Murray 2020; Liu and Kanoski 2018; Levine and Billington 1997]. This heightened reward value leads to the prioritisation of actions and behaviours to engage in food seeking and consummatory behaviour until interoceptive hunger signals are adequately addressed.

## 3.1 MiniGrid Environment

This study conducts a series of experiments using an 8x8 standard empty MiniGrid environment [Chevalier-Boisvert *et al.* 2018]. The standard environment contains 64 cells and 7 potential actions. In the standard environment, the agent only receives a reward for reaching the green goal square. This final reward is a function of the number of steps taken by the agent prior to reaching the goal state. This is formally defined as:

$$r = 1 - 0.9 \times \frac{\text{steps}_{\text{done}}}{\text{steps}_{\text{max}}}$$

- Where $\text{steps}_{max}$ is the maximum number of permissible steps before the episode truncates. This is set to 80 as detailed above.

The standard MiniGrid environment is reward scarce; meaning that an agent only receives a negative or positive reward at the end of the episode after either reaching the

green goal square or using the maximum number of steps permitted. This hampers the agent's learning as it relies on more immediate positive and negative feedback from the environment [Barto 2012]. This is why minor adjustments are made to the environment to facilitate improved learning and integrate food location within the grid. These are:

- Agent receives a fixed reward of +1 for reaching the goal state regardless of the number of steps taken prior.
- Agent receives a fixed reward of +1 for every step taken within the goal state (i.e. green goal square). This means the agent continues to get a reward for remaining at the goal state.
- Agent receives a fixed reward of -1 for every non-goal state transition. This includes bumping into walls.
- Agent can only take 3 possible actions: 0 for turning left, 1 for turning right and 2 for moving forward.

The standard MiniGrid environment returns state observations as an image of 7x7x3 array, direction and mission of agent; this is an incompatible input to the DQN agent. The environment is passed through a function to flatten this input into a single array to facilitate its input to the RL agent's DQN neural network. This function removes the walls of the 8x8 grid and returns a 6x6 grid shown in Figure 3.1. The agent is in the "fully observable" mode meaning that for each action it takes, it receives a next state observation which includes every possible state in the environment (i.e. the agent sees the entire grid world at every step). Each cell in the 6x6 grid is represented by 3 values:

- OBJECT_IDX – if the agent is in the cell.
- COLOR_IDX – the colour of the agent in the cell. This is set to red.
- STATE – additional information about the agent, such as its direction/orientation.

Thus, the agent receives an input of 108 (6x6x3) values representing the state of the entire environment (6x6 grid). At every state, or cell in the grid, the agent is aware of its position and direction shown in Figure 3.1. The DQN neural network is aware of the agent's position within the grid world and its orientation from the OBJECT_IDX and STATE information described above. However, the agent's position and direction are directly accessed in this study to embed a food location and food reward in the environment. At the start of every episode and after resetting the environment, the agent is located at the left corner of the 6x6 grid which corresponds to position (1,1) as shown in Figure 3.1.

## 3.2 Modelling Interoception

In existing RL pipelines, an agent learns an optimal policy to maximise external reward by attaining the goal state [Sutton *et al.* 1998]. Such RL agents do not perceive nor value interoceptive signals or intrinsic interoceptive rewards [Aubret *et al.* 2019; Chentanez *et al.* 2004]. Even RL agents which possess intrinsic motivation detailed in section

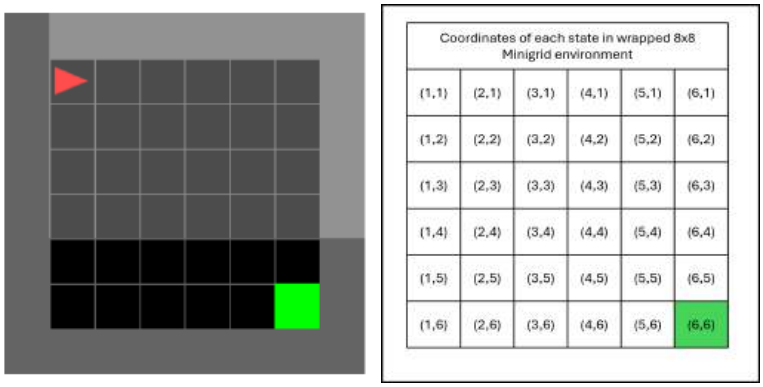



Figure 3.1: On the left, render image of wrapped 8x8 Empty Minigrid environment used for experiments. Red triangle represents agent and green square represents goal state. On the right, position coordinates of gridworld are represented as a coordinate of column and row.

2.3, do not have any motivation or reward incentive to maintain internal homeostasis. Hence, this study performs the dual process of representing interoceptive information to an RL agent and accurately reflecting its dynamic value during hunger and satiated states to the agent as experienced in living organisms.

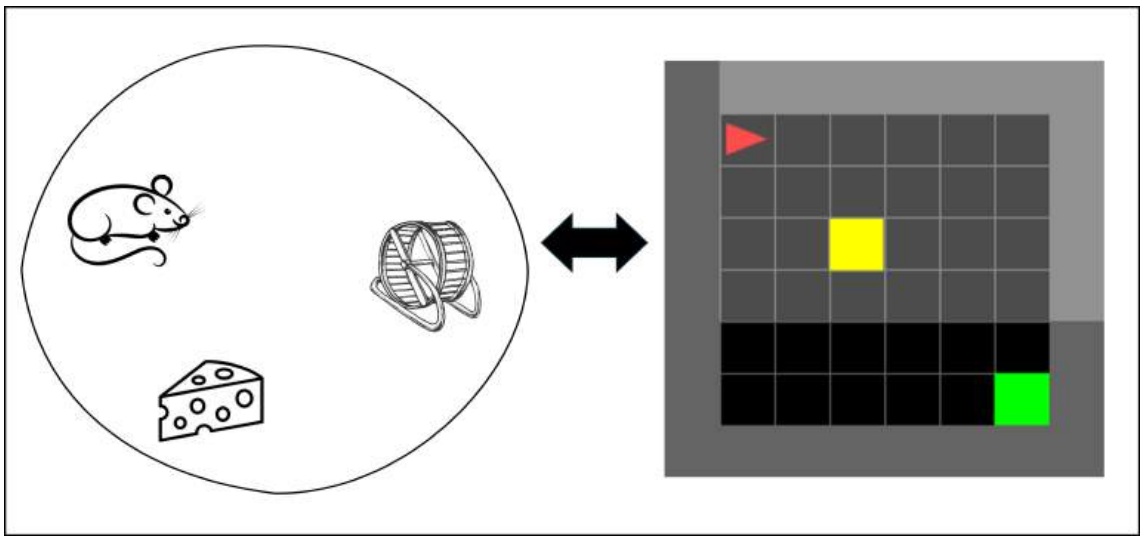


Figure 3.2: Analogy created by the experiment. From right-to-left, the red triangle RL agent represents the mouse. The yellow square is the food location and represents the block of cheese (intrinsically rewarding). The green square is the goal state of the environment and represents the running wheel (extrinsically rewarding).

This study recreates typical rodent experiments which investigate food seeking behaviour in mice. This is emulated by creating an analogy of a mouse in a flat room to an RL agent in a standard grid world environment as shown in Figure 3.2. Within this analogy, the RL agent represents the mouse, the goal state (i.e., the green square) represents a naturally pleasurable activity for a mouse, such as running on a wheel. The

mission of a standard RL agent in this environment is to reach the green square and its reward is derived from achieving this objective. The green square is the goal state of the environment and is extrinsically rewarding. RL agents without interoceptive information are able to learn a policy which maximises rewards from the goal state. The yellow square represents an accessible food location within the flat room, like the block of cheese in the analogy. This yellow square is intrinsically rewarding for the agent because it reduces ghrelin levels and is associated with an intrinsic food reward. The reward obtained from being in the yellow square is solely tied to the agent's ghrelin levels (i.e., internal state of hunger or context) and is not correlated to the objective of the environment of reaching the green goal square. Existing RL agents without interoceptive information do not form a policy to enter this state (i.e., yellow square) nor to maximise its reward.

To accurately reflect the mouse's perception of food in the flat room, it is necessary to associate food with a reward value perceived by the RL agent. This reflects the release of dopamine (i.e., reward) caused by eating food as detailed in section 2.5. Food reward must be a function of ghrelin levels and must be modulated depending on whether or not the agent is "hungry" to accurately reflect its heightened reward value detailed in section 2.5. The RL agent must receive ghrelin levels and possess predefined hunger thresholds. The RL agent receives ghrelin levels signals which mimic ghrelin secretion in the human body (section 2.5.1). Overtime and without eating, ghrelin levels naturally increase in living organisms. The experiment assumes a linear increase of ghrelin levels from 0 to 1. Ghrelin levels cannot be negative. At the start of each episode, ghrelin level is reset to 0 to recreate the effect of overnight fasting in rodent studies whereby the effect of prior meals on ghrelin levels is eliminated as detailed in section 2.4. This means the RL agent begins every episode "not hungry" or "fully satiated". In standard RL environments, an episode ends if the agent reaches the externally rewarding goal state (terminates) or if it takes the maximum number of permissible steps in the environment (truncates) [Sutton *et al.* 1998]. In all experiments, episodes run until truncation so the agent can take additional steps after reaching the goal state (green goal square). The maximum number of steps an agent can take is set to 80. This is fixed for all experiments. In conclusion, in every episode, the agent takes 80 steps and its ghrelin level linearly increases by 0.0125 from 0 to 1 over this duration as shown in Figure 3.3.

Different thresholds of ghrelin levels which categorise an agent as "hungry" were experimented with. These are referred to as hunger thresholds. Hunger threshold is a hyperparameter, set externally to reflect the amount of ghrelin that needs to be released in the gut in order to prompt actions from the brain in living organisms. The following experiments tested two hunger thresholds, 0.5 and 0.3, shown in Figure 3.4. Hunger threshold determines how quickly an agent gets hungry, and in turn affects the frequency of meals. Every action taken in the yellow square (i.e., food location) is considered an "eating instance" or meal. Experiments are designed such that the agent learns to keep its ghrelin levels below the hunger thresholds without any hardcoded requirements on how many meals should be taken by the agent; this is solely deter-

mined by the agent's ghrelin level and the ghrelin decrease caused by the food location detailed below.

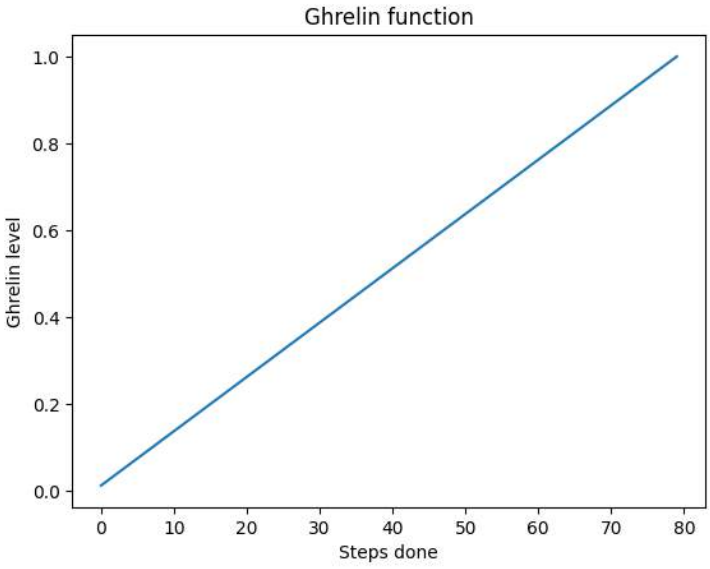



Figure 3.3: Modeled Ghrelin function, which linearly increases from 0 to 1 over 80 permissible steps.

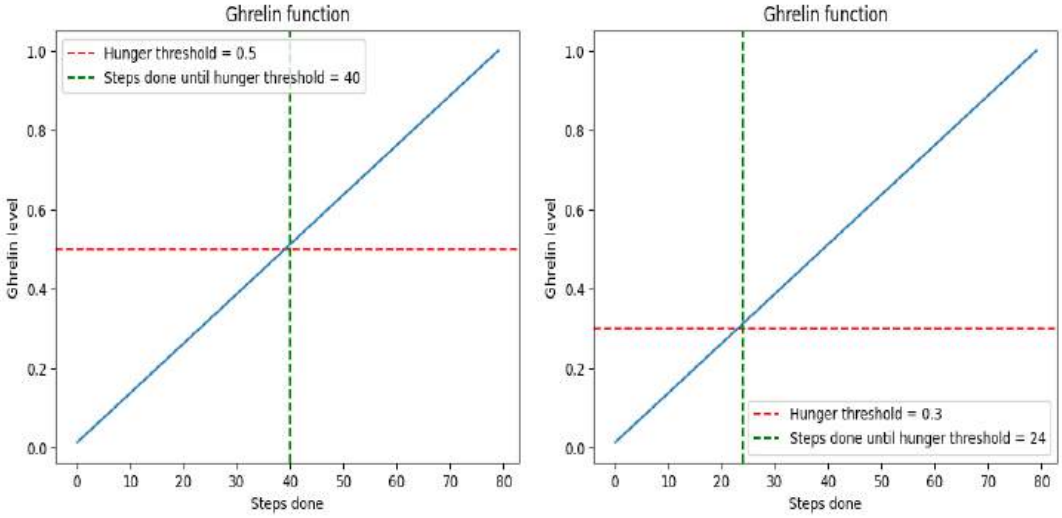



Figure 3.4: Illustration on the left of hunger threshold = 0.5 and on the right of hunger threshold = 0.3

After modelling interoceptive hunger signals through incremental ghrelin levels and set hunger thresholds, food and its impact on ghrelin levels (i.e., decreases ghrelin levels and causes DA release in brain) needs to also be reflected in terms understood by the RL agent. Food location and food reward are incorporated to achieve this respectively. Food location (i.e., yellow square in Figure 3.2) is a position or state within the Mini-Grid environment for which the agent receives a positive food reward for being in and

a negative food reward for being in any other position within the grid (i.e., similar to the reward structure for the goal state in the standard MiniGrid environment detailed in section 3.1). Incorporating ghrelin levels and food reward serves a two-fold purpose: (1) ghrelin levels are added to state information as context. This allows the agent to associate which ghrelin levels (i.e., hunger thresholds) result in negative food reward.; (2) food location and food reward mirror the biological effect of food on lowering ghrelin levels and its rewarding nature respectively. Food reward information is separately stored in the DQN agent's replay memory to associate state transitions with food reward obtained. This enables the agent to learn that being at the food location (i.e., "eating") is pleasurable.

The effect of food (i.e., agent "eating" at the food location) on ghrelin levels is modeled as a decrease of 0.1 for every step the agent takes at the designated food location. This value of 0.1 was chosen for its ratio to the minimum and maximum value of ghrelin modeled (0 and 1 respectively). This impact is fixed across all experiments. To reflect the heightened perceived value of food when the agent is hungry versus satiated based on the synthesised model detailed in section 2.5, every experiment contains two possible values of food reward: a baseline food reward and a heightened food reward value. The baseline food reward is a small reward value that the agent receives for being at the food location (i.e., "eating") when it is not "hungry" (i.e., its ghrelin levels are below the hunger threshold). This is modeled as +0.5 for being in the food location and -0.5 for being elsewhere on the grid when the agent is not "hungry" shown in Figure 3.5. This baseline food reward is fixed across all experiments. The selection of 0.5 is to reflect the sub-par rewarding nature of food in comparison to other activities when an organism is not hungry based on conclusions from the literature detailed in section 2.5. In practice, +0.5/-0.5 is numerically lower than the reward for reaching the green goal square (i.e., +1/-1) of the MiniGrid environment.

The heightened food reward is a large food reward value that the agent receives for being at the food location within the MiniGrid above its hunger threshold (i.e., the value of "eating" when an agent is "hungry"). This is a high positive food reward for being at the food location above the hunger threshold and a high negative food reward for being elsewhere. Experiments detailed in section 3.5.2 test with varying values of this heightened food reward value to emulate the comprehensive evaluation of food-seeking behaviours at the OFC-Amygdala-NAcc subsystem as single unit. Based on conclusions from section 2.5, the OFC is part of the PFC and is also part of the OFC-Amygdala-NAcc subsystem which conducts a comprehensive evaluation process and sends a decision to other PFC areas (dlPFC, mPFC and ACC) to develop and execute food-seeking plans illustrated in Figure 2.6. The above set up caters to the observed value of food and food seeking behaviour by the OFC-Amygdala-NAcc subsystem when the agent is hungry versus satiated while maintaining the physiological effect of food across both scenarios. Different combinations of hunger thresholds, food locations and heightened food reward are experimented in section 3.5 to better demonstrate the role of the PFC in regulating food intake. Although the implementation of baseline and heightened food reward is motivated by the literature detailed in section 2.5, alterna-

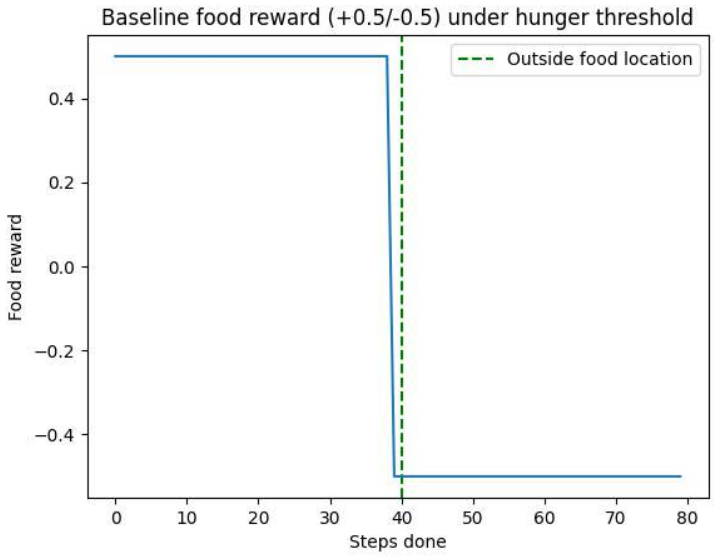



Figure 3.5: Baseline food reward under hunger threshold. Agent receives +0.5 for steps done in food location and -0.5 for any step elsewhere on the grid as described in section 3.2

tive experiments which implement a proportional food reward function and internal ghrelin controlling actions are explored in section 4.3.1.

## 3.3 Integrating interoceptive intrinsic reward into DQN

Building on conclusions by Latyshev and Panov [2024]; Kayal *et al.* [2025] described in section 2.3, this study integrates intrinsic motivation through the use of a complementary intrinsic reward signal. Unlike previous work by Barto *et al.* [2004] which integrates intrinsic reward through a combination of Q-learning and semi-Markov Decision Process (SMDP), this study instead utilises a CMDP and Q-learning which integrates a total reward signal (i.e., both extrinsic and intrinsic reward) in the calculation of state-action Q-values. As noted in section 2.2, a DQN agent aims to maximise discounted future reward by selecting maximum action-values (Q(s,a)) for all future states. This obeys the Bellman equation equation of state-action Q values defined as:

$$Q^{\pi}(s,a) = r + \gamma Q^{\pi}\left(s', \pi(s')\right)$$

Intrinsic reward complements extrinsic reward as described in section 2.3 using the equation:

$$r_{\text{total}} = r_{\text{ext}} + \beta * r_{\text{int}}$$

where:

- $r_{\text{ext}}$ – extrinsic reward
- $\beta$ – intrinsic reward coefficient
- $r_{\text{in}}$ – intrinsic reward

This adjusted total reward signal is used in the calculation of predicted next state-action Q-values, to give the adjusted equation below:

$$Q^{\pi}(s,a) = (r_e + \beta * r_i) + \gamma Q^{\pi}\left(s', \pi(s')\right)$$

Where:

- $r_e$ – Extrinsic episodic reward achieved by the agent for reaching the goal square or otherwise (either +1/-1)
- $r_i$ – Intrinsic reward achieved by the agent for reaching food location or otherwise (either +0.5/-0.5 under the hunger threshold and different values detailed in section 3.5.2)
- $\beta$ - intrinsic reward coefficient which is set to 1

The intrinsic reward coefficient is a hyperparameter and is set to 1 in this study. This is because DA released in the human brain by actions which address interoceptive needs like eating, compared to DA released by external rewards activates the same neural pathways. In effect, regardless of the motivation of the reward (i.e., intrinsic versus extrinsic), dopamine released in the brain activates the same mesolimbic pathway in the brain [Baik 2021; Chau *et al.* 2018]. Furthermore, other implementations of intrinsic motivation and reward utilise intrinsic reward functions that are proxies of exploration, such as information gain or state count, and thus, incorporate a decaying or fractional intrinsic reward coefficient [Latyshev and Panov 2024; Kayal *et al.* 2025]. These reward functions are not applicable to this study's use of intrinsic reward as the intrinsic reward signal is only dependent on the RL agent's ghrelin levels with no correlation to its performance on the base environment's mission (i.e., exploring enough to reach the green goal square).

## 3.4 Architecture and Hyperparameters

The experiment utilised a fully connected neural network to perform deep Q-learning. This architecture contains an input layer with input nodes that match the flattened number of observations of the 8x8 Minigrid environment. It includes 2 hidden layers of 128 nodes, and an output layer of 3 nodes representing the three possible actions (turning left, turning right, moving forward). The experiments test two settings of interoceptive information: (1) complete omission and (2) incorporation as context. To incorporate interoceptive information (ghrelin levels) to the input of the RL agent, it was concatenated to state information, following a similar method used by Ball *et al.* [2021]. This is an additional input node to the input layer of the DQN representing ghrelin levels shown in Figure 3.6.

As detailed in chapter 2, the agent learns to minimise the error between predicted next state-action Q-values and sampled state-action Q-values from the environment. The following experiments use the mean squared error (MSE) to compute this error from samples in the replay memory and gradient descent (GD) to backpropagate through the DQN network to update its weights. GD computes the gradient of the loss function

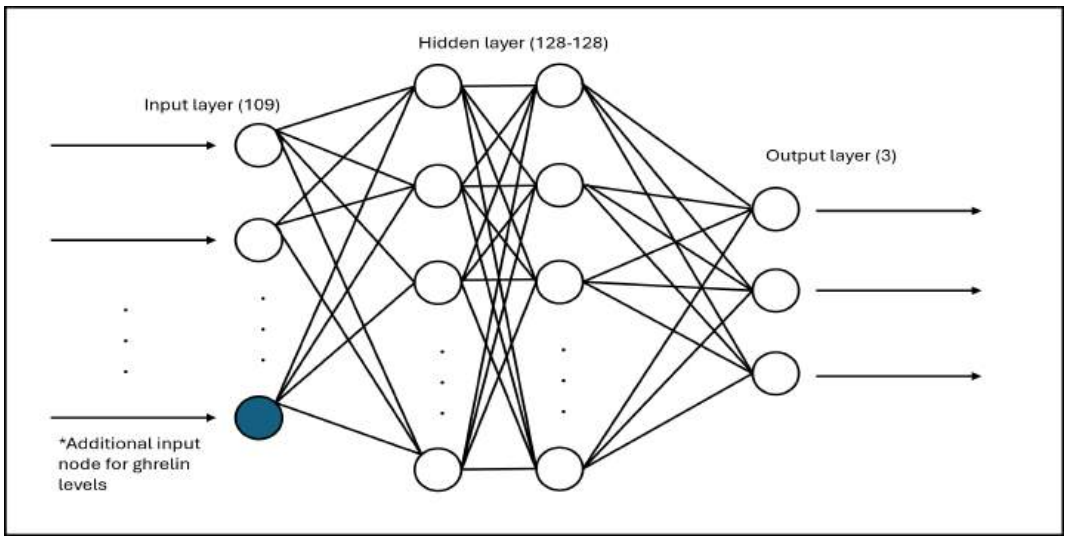



Figure 3.6: Architecture of DQN neural network used by agent. DQN receives state information and ghrelin levels (109 nodes) and outputs an action (3 nodes).

with respect to the parameters of the node weights for the entire training dataset (i.e., samples in the replay memory) [Ruder 2016]. This ensures that the addition of an input node for ghrelin levels that is constantly 0 as described in section 3.5 does not influence the network's ability to learn the pre-existing relationship between states and actions in the environment. This influence is otherwise present when ghrelin levels are not 0. To facilitate feature learning, the DQN network weights are randomly initialised from a Gaussian distribution with mean of 0 and standard deviation of 0.01 [Narkhede *et al.* 2022]. Additional hyperparameters are detailed in Annex A.

## 3.5 Summary of Experiment Set-ups

### 3.5.1 Baseline Experiments with Interoception Omitted

#### Standard RL agent

This experiment deploys a standard RL agent in an 8x8 MiniGrid environment (detailed in section 3.1). This agent does not possess any ghrelin levels, nor receives any food reward. There is no embedded food location within this experiment's environment. The environment reward detailed in section 3.1 is the same and every episode runs until it truncates. The agent continues to receive a positive reward of +1 for remaining at the green goal square (goal state). This experiment is standard RL and does not require further validation. Its purpose is to demonstrate that the selected hyperparameters enable the RL agent to solve the environment's task and to serve as a baseline of the agent's behaviour for subsequent experiments.

#### RL agent with static interoception (zero ghrelin)

This set up involves the same agent detailed above, except the agent receives an additional input of its ghrelin levels which is constantly 0. The same settings detailed above apply and the environment has no food reward or embedded food location. Referring to the analogy detailed in Figure 3.2 and literature in section 2.5, this experiment is parallel to disrupting the nucleus solitary tract (NTS) that conveys the secretion of ghrelin in the stomach to the brain. Without this vital relay connection to the brain, it is equivalent to not having any ghrelin secreted in the stomach (i.e., ghrelin equals zero). This experiment is also to ensure that adding an additional empty input to the DQN agent does not affect its learning.

#### RL agent with interoceptive signals but no food reward

This experiment tests the incorporation of interoceptive signals (ghrelin levels, hunger threshold and food location) without the agent receiving any food reward. This experiment deploys a standard RL agent with an additional input conveying its linearly increasing ghrelin levels. There is an embedded food location within the grid world which causes ghrelin levels to decrease as detailed previously. The hunger threshold is set to 0.5 and food location is set to position (3,3) within the grid as shown in Figure 3.8. The agent does not receive any food reward for visiting this location even though it possesses a hunger threshold (i.e. it becomes “hungry”). Referring to the analogy detailed in Figure 3.2 and literature in section 2.5, this experiment is equivalent to placing a mouse in a flat room with access to inedible objects or placebo pellets. The mouse does get hungry (i.e., linearly increasing ghrelin levels which exceeds hunger threshold) but because the objects are not real food, they do not cause a release of dopamine in the brain (i.e., the agent does not experience a food reward) even if the mouse eats them (i.e., if the agent intersects the food location). Although the reward (i.e., dopamine release) associated with eating food is intrinsic, it is only obtained when the organism takes the external action of eating similar to the RL agent taking the external action of visiting the food location. However, without the potential to ”die” from hunger, the RL agent is not fully incorporating interoceptive signals related to hunger. This experiment further highlights to that accurately recreate the analogy in Figure 3.2, the RL agent must receive a positive reward which represents food’s rewarding nature and also a negative reward which represents the pain associated with hunger (i.e., not eating or visiting the food location).

### 3.5.2 Experiments with Interoception fully incorporated

Experiments which integrate interoception are designed by manipulating the three key parameters - hunger threshold, heightened food reward and food location. These key parameters are explained in section 3.5.2 and section 3.5.2 before elaborating on the different control experiment set ups.

### Manipulating value of food reward

As noted in Chapter 3, the value of food reward depends on whether the agent is ”satiated” (ghrelin level below hunger threshold) or ”hungry” (ghrelin level above hunger threshold). When the agent is ”satiated”, it receives a baseline food reward of +0.5 and -0.5 as shown in Figure 3.5. Heightened food reward is a key aspect of modelling interoception in an RL pipeline as this value must reflect the holistic evaluation of food seeking behaviour conducted by the OFC-Amygdala-NAcc subsytem when an organism is hungry versus satiated. Experiments test different values of heightened food reward which the agent receives for being at a food location when its hunger threshold is exceeded versus any other state or position within the grid.

1. The agent receives a reward of +1 for steps taken in the food location when “hungry” and receives a reward of -1 for any other location/state within the grid shown on the left of Figure 3.7.

2. The agent receives a reward of +2 for steps taken in the food location when “hungry” and receives a reward of -2 for any other location/state within the grid shown on the center of Figure 3.7.

3. The agent receives a reward of +5 for steps taken in the food location when “hungry” and receives a reward of -5 for any other location/state within the grid shown on the right of Figure 3.7.

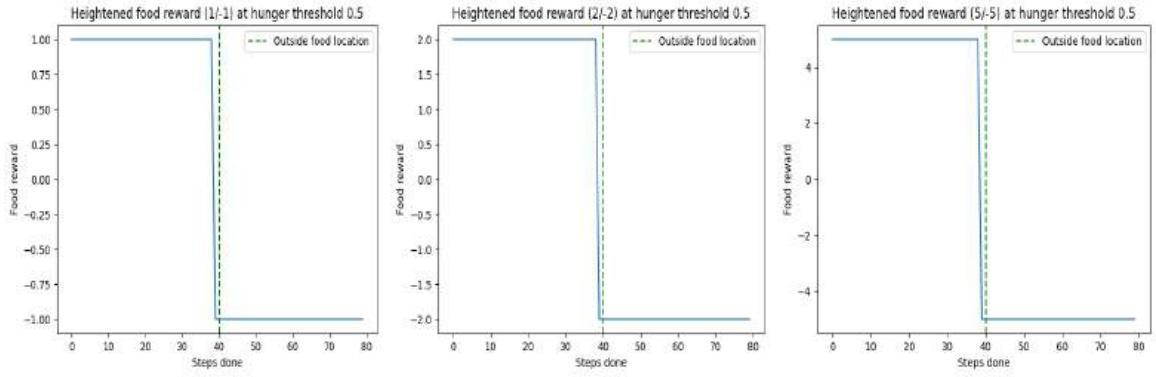



Figure 3.7: From left-to-right, graphical representation of heightened food reward of (+1/-1), (+2/-2) and (+5/-5) when agent is deployed ”hungry” (with static ghrelin level = hunger threshold = 0.5) without any embedded food location.

### Manipulating food location

To investigate the influence of interoceptive information on an agent’s learned policy, different positions within the grid are set as the food location. These were two main kinds of locations:

1. Food location is placed at a significant distance from the green goal state. These experiments test with food location (3,3) as shown in Figure 3.8 and (4,2) as shown in Figure 3.9 on the 8x8 MiniGrid environment.

2. Food location is placed at a closer distance to the green goal square (i.e. goal state of environment). This experiment sets food location to (5,4) as shown in Figure 3.10.

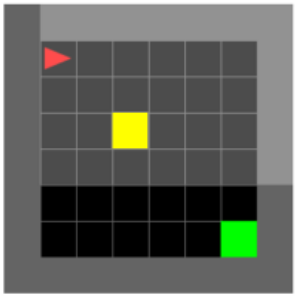


Figure 3.8: View of 6x6 grid world environment when food location (represented as yellow square) is set to position (3,3) within the grid.

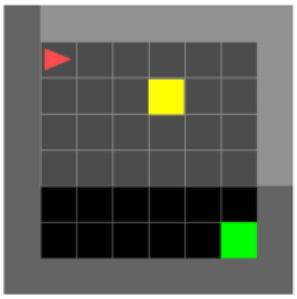


Figure 3.9: View of 6x6 grid world environment when food location (represented as yellow square) is set to (4,2) within the grid.

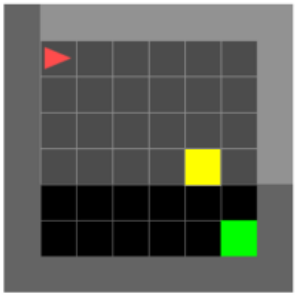


Figure 3.10: View of 6x6 MiniGrid when food location (represented as yellow square) is set to (5,4) within the grid.

A series of experiments controlling for one of the three key variables – hunger threshold, food location and heightened food reward – are conducted as follows:

- To investigate the influence of heightened food reward by using a fixed hunger threshold and food location but different heightened food reward values. In these experiments, the hunger threshold is fixed at 0.5 and food location is set to position (3,3) within the grid as shown in Figure 3.8. Heightened food reward of +1/-1, +2/-2 and +5/-5 are tested as shown in Figure 3.7.

- To investigate the influence of food location by using a fixed hunger threshold and heightened food reward but different food locations. In these experiments, the hunger threshold is set to 0.5 and the three values of heightened food reward are each tested with food located at positions (4,2) and (5,4) as shown in Figure 3.9 and Figure 3.10 respectively.
- To investigate the influence of hunger thresholds by using a fixed heightened reward value and food location but different hunger thresholds. These experiments repeat all the above with hunger threshold set to 0.3 as shown in Figure 3.4.

# Chapter 4

# Results

## 4.1 Baseline experiments with interoception omitted

### 4.1.1 Standard RL agent

The first baseline experiment is a standard RL agent deployed to solve the 8x8 MiniGrid environment (described in detail in section 3.5). This standard RL agent learns with the same parameters, conditions and optimisation algorithms in Chapter 3 without ghrelin levels, food reward or food location present in the grid world. The agent learns an optimal policy to reach the green square and the learned path is illustrated in Figure 4.1. This same RL agent is deployed but receives an additional input of ghrelin levels that is kept constant at 0 as detailed in shown in Figure 4.2. This RL agent also learns the same policy shown in Figure 4.1.

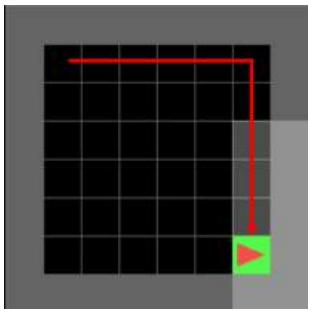


Figure 4.1: Illustration of learned policy by standard RL agent without interoceptive information described in section 4.1.1.

### 4.1.2 RL agent with interoceptive signals but no food reward

In this experiment, the previous RL agent with interoceptive signals is deployed except with linearly increasing ghrelin levels as shown in Figure 3.3. This agent possesses a hunger threshold of 0.5 as shown in Figure 3.4 but this threshold is associated with a static food reward of 0. A designated food location within the grid is at position

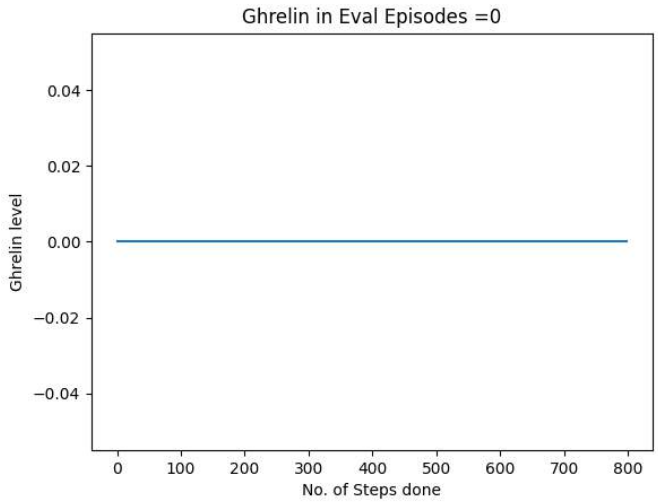



Figure 4.2: Graph showing static 0 ghrelin levels sent to RL agent described in section 4.1.1.

(3,3) and causes a decrease in ghrelin levels by 0.1 but the agent does not receive any food reward for visiting this state within the grid world (i.e., food reward is constantly 0). This is to recreate the parallel detailed in section 3.5.1 of a mouse placed in a flat room with an inedible object. The agent learns the same optimal policy as the previous baseline settings illustrated in Figure 4.3. The impact of food location on ghrelin levels is never seen during evaluation because the agent learns a policy that does not intersect this location. This shows that despite receiving interoceptive information, without any food reward feedback, the RL agent is unable to learn the correlation between the food location and ghrelin levels. In essence, the agent never learns the value of "eating" and thus, never engages in it (i.e., by visiting the food location (3,3)). This behaviour is in line with fundamentals of RL and does not require further validation [Sutton *et al.* 1998]. However, its inclusion in this section is to demonstrate the RL agent's learned path for later comparison with additional experiments detailed in section 4.3 where the agent has actions to internally control its ghrelin levels.

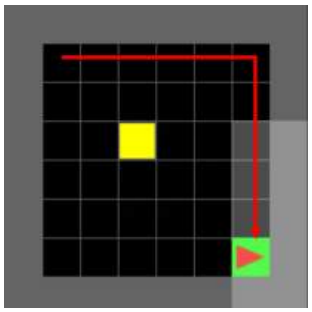


Figure 4.3: Illustration of the grid world environment in experiment where RL agent possess interoceptive signals but receives a constant food reward of 0. Red line demonstrates trajectory taken by agent. Yellow square represents food location with the grid world at position (3,3).

## 4.2 Experiments with Interoception fully incorporated

### 4.2.1 Experiments investigating the influence of heightened food reward

The agent receives two values of food reward depending on whether it is above or below its hunger threshold ("hungry" versus "satiated") as detailed in Chapter 3. Thus, the perceived value of food changes between these two states, which this study labels as baseline food reward (when agent is "satiated") and heightened food reward (when agent is "hungry").

The following experiments are set to a hunger threshold of 0.5 and food is located at position (3,3) within the grid world shown in Figure 3.4 and Figure 3.8 respectively. The baseline food reward value is fixed and set to +0.5 and -0.5 when the agent is below the hunger threshold. Three values of heightened food reward are tested: +1 and -1; +2 and -2; +5 and -5. As shown in Figure 3.4, when the hunger threshold is 0.5, the agent becomes "hungry" at time step 40. After 40 steps, the agent begins to receive this heightened food reward value.

Across all three heightened food reward values, the RL agent learns a policy which intersects the food located at (3,3) enroute to the green goal square illustrated in Figure 4.4. The key difference is observed when the agent receives a heightened food reward of +1/-1 compared to +2/-2 and +5/-5. When heightened food reward is +1/-1, the agent remains at the green goal square despite exceeding its hunger threshold of 0.5. In the other two settings of heightened food reward, the RL agent leaves the green goal square to visit the food location when it becomes "hungry". It then stays at the food location until its ghrelin levels drop back to 0. This is interesting because a single step within the food location is enough to effectively return the agent to a "satiated" state (i.e., ghrelin levels below its hunger threshold). The agent independently learns to remain at the food location until its ghrelin levels return to 0. After which, it returns to the green goal square. This trajectory is shown in Figure 4.5. There is also a difference between the agent's behaviour when heightened food reward is +2/-2 and +5/-5 despite visiting the food location in both settings. When heightened food reward is +2/-2, the agent visits the food location after its hunger threshold is exceeded as shown in Figure 4.9. In contrast, when heightened food reward is +5/-5, the agent visits the food location before exceeding its hunger threshold (0.5) as shown in Figure 4.10.

The behaviour of the RL agent across different settings of heightened food reward are interpretable when the total episode reward is considered shown in Figure 4.6, Figure 4.7 and Figure 4.8. Total episode reward is the sum of food reward and episode reward, defined below:

$$r_{tr} = r_{env} + r_{fr}$$

where:

- $r_{env}$ is the extrinsic reward for being at the green goal square

- $r_{fr}$ is the intrinsic food reward for being at the food location
- $r_{tr}$ is total episode reward

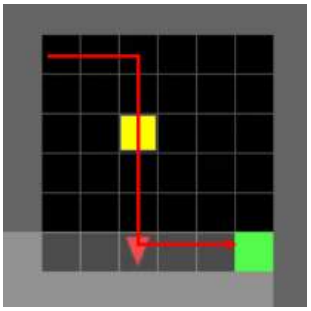


Figure 4.4: Learned policy by RL agents with heightened food reward at hunger threshold 0.5 as described in section 4.2.1

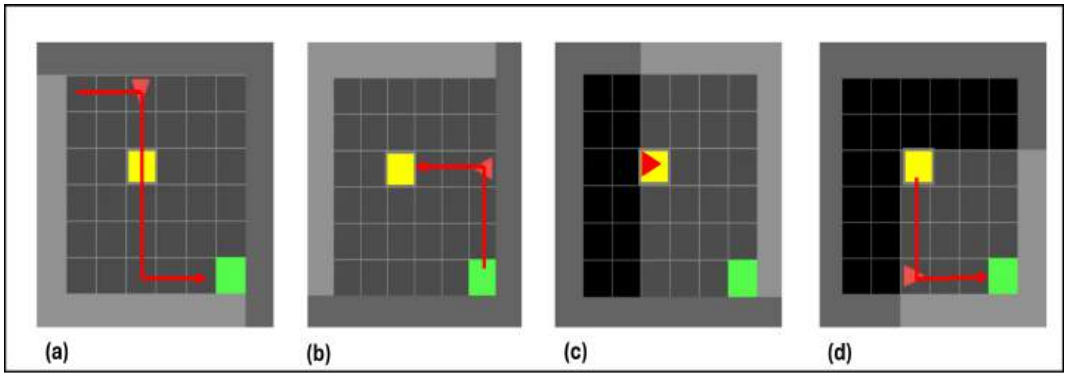



Figure 4.5: Learned policy of RL agent which receives a heightened food reward of +5/-5 above hunger threshold 0.5. Yellow square represents food located at position (3,3). From left to right, (a) Red line shows the trajectory of the agent at the start of the evaluation episode which passes food location (3,3). (b) After some steps at the goal state (6,6), the agent leaves the green goal square in search of food as ghrelin levels near the hunger threshold 0.5. Agent officially becomes "hungry" (c) The agent enters the food location (3,3) and takes approximately 5 steps in the food location which result in a decrease of ghrelin levels by 0.5, returning it to 0. The agent is now "fully satiated". (d) Afterwards, the agent returns to the goal state where it remains until the episode truncates. The agent behaves similarly when it receives a heightened food reward of +2/-2.

When heightened food reward is +1/-1 (Figure 4.6), the RL agent remains at the green goal square where it receives a reward of +1 for being in this goal state even when the agent becomes "hungry". The total episode is reward is 0 (not negative) and explains why the agent is not incentivised to leave the goal square to reach the food location.

$$1_{env} + (-1_{fr}) = 0_{tr}$$

Both heightened food reward of +2/-2 and +5/-5 are above the base environment's reward of +1/-1. This explains why the agent is motivated to leave the green goal

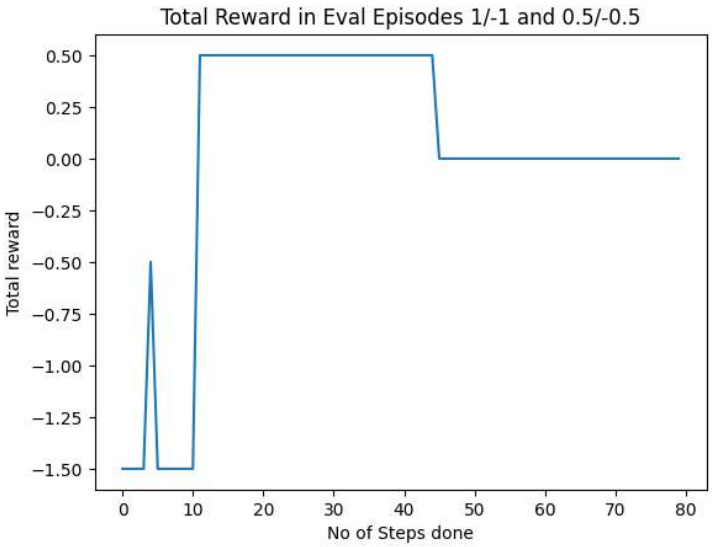



Figure 4.6: Total episode reward when RL agent receives a heightened food reward of +1/-1 when hunger thresholds is 0.5.

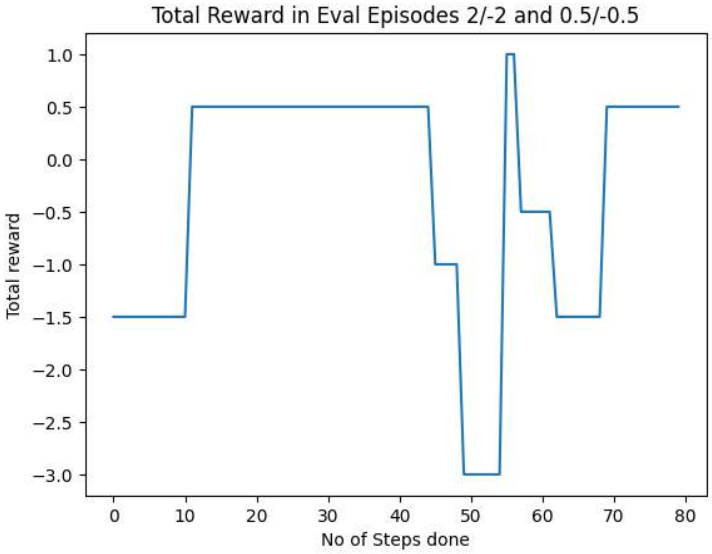



Figure 4.7: Total reward received by RL agent when heightened food reward is +2/-2 and hunger threshold is 0.5.

square when its hunger threshold is exceeded to revisit the food location as shown by total episode reward in Figure 4.7 and Figure 4.8.

The step and ghrelin level at which the RL agent leaves the green goal square to the food location are different when heightened food reward is double the base environment's reward (i.e., +2/-2) compared to when it is five-times that value (i.e., +5/-5). When heightened food reward is +5/-5, the agent instinctively leaves the green goal square at step 33 (7 steps till threshold of 0.5 at step 40 is reached) to arrive at the food location before reaching its hunger threshold. As shown in Figure 4.8, the agent

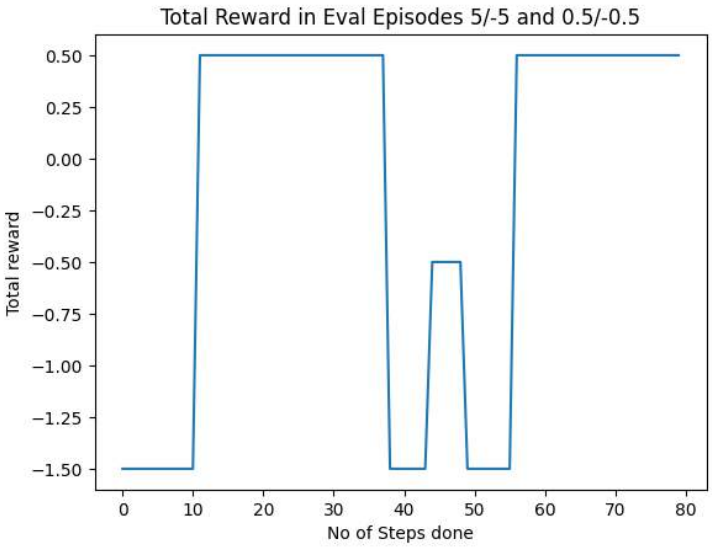



Figure 4.8: Total reward received by RL agent when heightened food reward is +5/-5 and hunger threshold is 0.5.

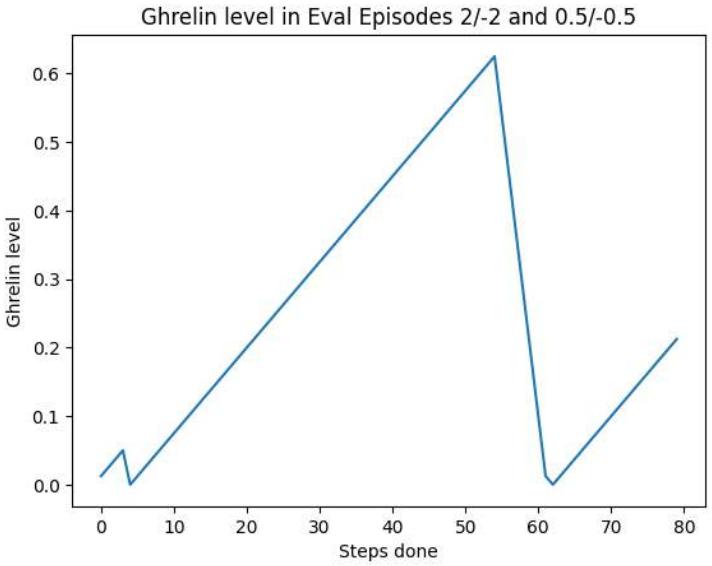



Figure 4.9: Ghrelin levels of RL agent when heightened food reward is+2/-2 at hunger threshold 0.5

only receives a reward of +0.5 for being at the food location (because it is still "satiated") when it reaches the location, bringing the total reward for steps within the food location:

$$-1_{env} + 0.5_{fr} = -0.5_{tr}$$

It remains at the food location for 5 steps, seen in the decline of ghrelin levels from 0.5 to 0 between step 40 and 48 shown in Figure 4.10. Once the agent decreases its ghrelin levels back to 0, it leaves the food location to return to the green goal square (between step 48 and 56). When it returns to the green goal square, the total reward shown in

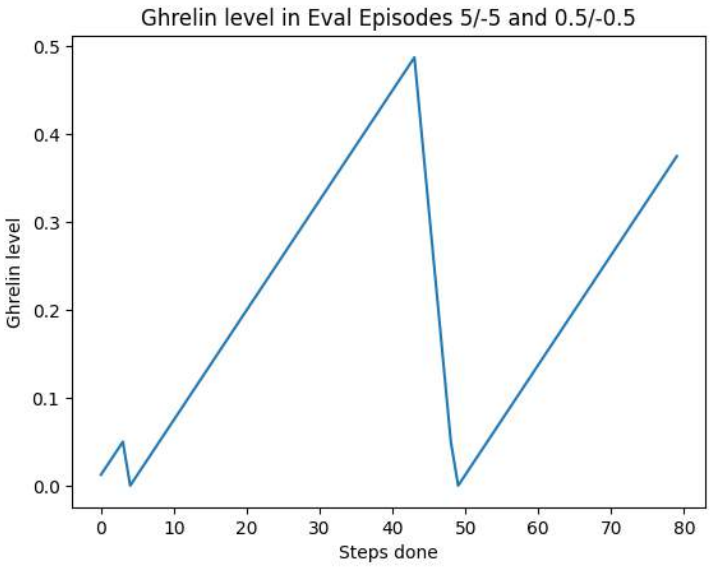



Figure 4.10: Ghrelin levels of RL agent when heightened food reward is +5/-5 at hunger threshold 0.5

Figure 4.8 is static at 0.5:

$$1_{env} + (-0.5_{fr}) = 0.5_{tr}$$

When heightened food reward is +2/-2, the agent leaves the green goal square at approximately step 45 (5 steps after exceeding its threshold of 0.5 at step 40) as shown in Figure 4.7 and Figure 4.9. This means it is already "hungry" (i.e., it receives the heightened negative food reward of -2) when it revisits the food location. The total reward between step 49 and 54 is:

$$-1_{env} + (-2_{fr}) = -3_{tr}$$

The agent stays approximately 7 steps in the food location, lowering its ghrelin level back to 0 until it shifts from "hungry" to "fully satiated" shown in Figure 4.9 and food reward switches back from +2 to +0.5 for being at the food location. The RL agent then follows the same trajectory as when heightened food reward is +5/-5 and returns to the green goal state illustrated in Figure 4.5. The agent then remains at the green goal square until the episode truncates, receiving a static reward of 0.5:

$$1_{env} + (-0.5_{fr}) = 0.5_{tr}$$

In conclusion, heightened food reward enables the RL agent to learn an optimal policy which adequately addresses its interoceptive needs alongside externally rewarding tasks. This is demonstrated by the ability of an RL agent with heightened food reward of +2/-2 and +5/-5 to lower its ghrelin levels by revisiting the food location after reaching the green goal square unlike the RL agent that receives a heightened food reward of +1/-1. An RL agent's ability to manage its ghrelin levels (interoceptive needs) is also affected by the value of the heightened food reward. This is demonstrated by the pre-emptive behaviour of the RL agent to visit the food location before exceeding

its hunger threshold (shown in Figure 4.8) when heightened food reward was +5/-5 (five-times the base environment's reward) compared to when heightened food reward was +2/-2.

### 4.2.2 Experiments investigating the influence of food location

To investigate the influence of food location on an RL agent's learned policy, the RL agent was deployed in environments where food location was embedded in different positions on the grid as detailed in Chapter 3. In these experiments, hunger threshold is fixed at 0.5. An RL agent with heightened food reward of +1/-1 is deployed in the grid world where food location is set to position (4,2). The agent learns a policy which intersects this location enroute to the green goal square shown in Figure 4.11. It remains in the green goal square until the episode truncates. This is the same behaviour demonstrated in section 4.2.1.

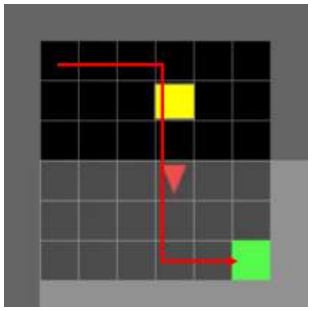


Figure 4.11: Red line illustrates learned path taken by RL agent when heightened food reward +1/-1 and hunger threshold 0.5. Yellow square represented food located at (4,2).

When the RL agent receives a heightened reward +2/-2 and +5/-5, and food is located at (4,2), it learns a policy which intersects the food location enroute to the green goal square as shown in Figure 4.11. The agent then behaves in the same manner as illustrated in Figure 4.5. In both cases, it revisits this food location to reduce its ghrelin levels back to 0 and then returns to the green goal square.

When an RL agent receives a heightened food reward of +1/-1 is deployed in gri- world where food location is at (5,4), the agent is able to manage its hunger levels by revisiting the food location as shown in Figure 4.12. This suggests that the proximity of food location (intrinsic reward) and green goal square (extrinsic reward) contributes to the agent's ability to manage its interoceptive needs. This is further examined in additional experiments detailed in section 4.3.1 and section 4.3.6.

An RL agent with heightened food reward of +2/-2 is deployed in the grid world where food location is set to position (5,4) as shown in Figure 3.10. The agent learns a policy which intersects this location enroute to the green goal square. It then leaves the green goal square to visit the food location. It remains at the food location until its ghrelin levels drop back to 0 and then returns to the green goal square. This trajectory is

illustrated in Figure 4.13. This is the same kind of behaviour exhibited by the agent shown in Figure 4.5.

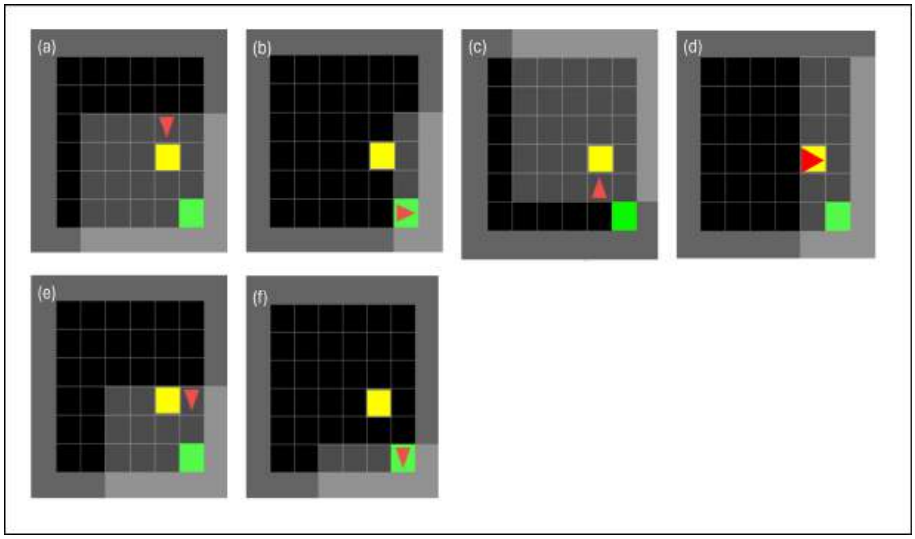



Figure 4.12: Frames of path taken by RL agent when heightened food reward +1/-1 and hunger threshold 0.5. Yellow square represented food located at (5,4).

When the RL agent receives a heightened food reward of +5/-5, it leaves the goal square at step 37 (3 steps before the hunger threshold 0.5 is reached) illustrated in Figure 4.15. The agent is able to reach the food location when it becomes "hungry" because this food location is close to the green goal square. When the agent receives a heightened food reward of +2/-2, it behaves the same way shown in Figure 4.14; leaving the green goal square just in time to reach the food location when it first becomes "hungry". This explains why the total reward plots of RL agents deployed with heightened food reward +5/-5 and +2/-2 are very similar shown in Figure 4.15 and Figure 4.9 respectively.

This is noticeably more controlled behaviour by the RL agent which receives a heightened food reward of +2/-2 with food located nearby in comparison to when food location was (3,3). As shown in Figure 4.9, the RL agent exceeded its threshold level before reaching the food location. In conclusion, food location affects the learned policy of the agent because the agent ensures that it intersects this food location (wherever it is located) enroute to the green goal square. The proximity of the food location plays a key role in enabling an RL agent manage its ghrelin levels regardless of heightened food reward. Across all heightened food reward values, the RL agent is able to effectively manage its ghrelin levels when the food location is close to the green goal square (i.e., food located at (5,4)) With extreme heightened food reward (+2/-2 and +5/-5), a proximal food location affects the agent's learned policy in maintaining its hunger thresholds, by teaching it when to leave the green goal square before reaching its hunger threshold. As shown in Figure 4.14 and Figure 4.15, this enables the agent to reach the food location at the exact time step when it effectively becomes "hungry" without experiencing extreme negative food reward.

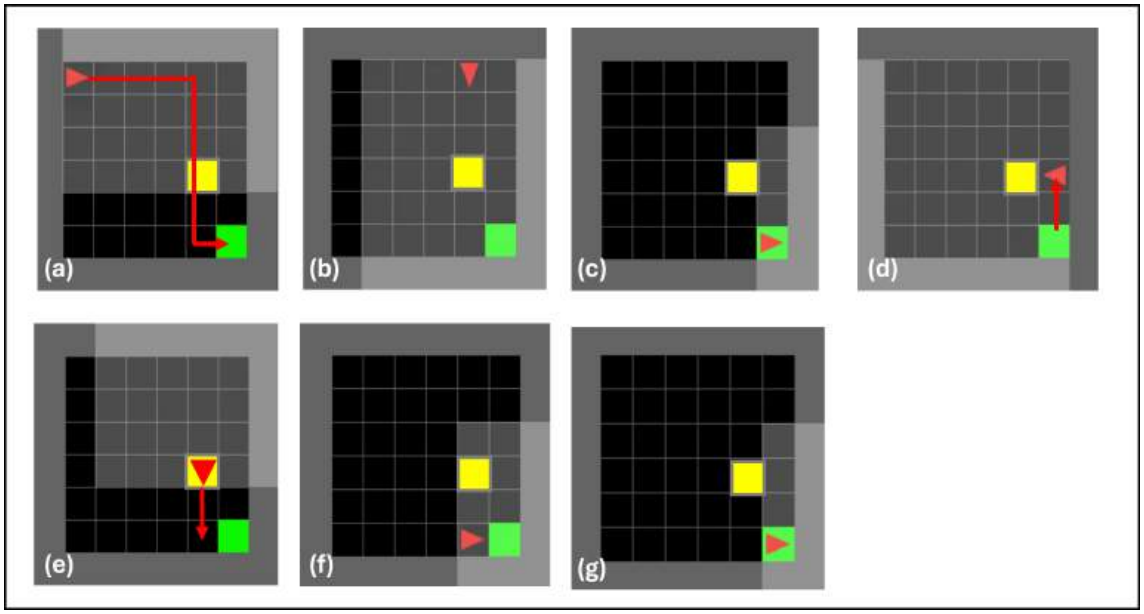



Figure 4.13: Learned policy of RL agent which receives a heightened food reward of +2/-2 above hunger threshold 0.5. Yellow square represents food located at position (5,4). From top left to bottom right, (a) Red line shows the trajectory of the agent at the start of the evaluation episode. (b) this trajectory passes food location (5,4). (c) The agent reached the goal state (6,6). (d) After some steps at the goal state (6,6), the agent leaves the green goal square in search of food as ghrelin levels reach the hunger threshold 0.5. Agent is "hungry" (e) The agent enters the food location (5,4) and takes approximately 5 steps in the food location which result in a decrease of ghrelin levels by 0.5, returning it to 0. The agent is now "fully satiated" again. (f) The agent leaves the food location to return to the goal state. (g) it remains at the goal state until the episode truncates. The agent behaves similarly when it receives a heightened food reward of +5/-5.

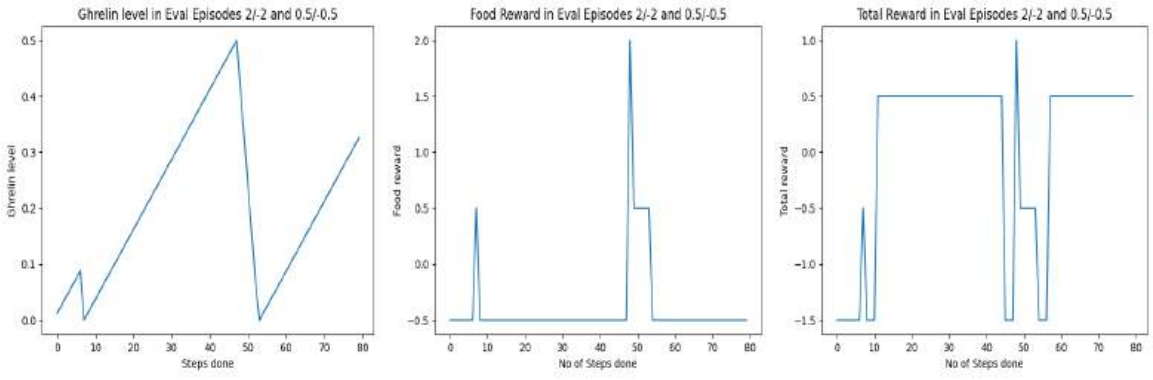



Figure 4.14: On the right, ghrelin levels of RL agent with heightened food reward +2/-2, food located at (5,4) and hunger threshold is 0.5. In the center, food reward received by the same RL agent. On the right, total reward received by the same RL agent.

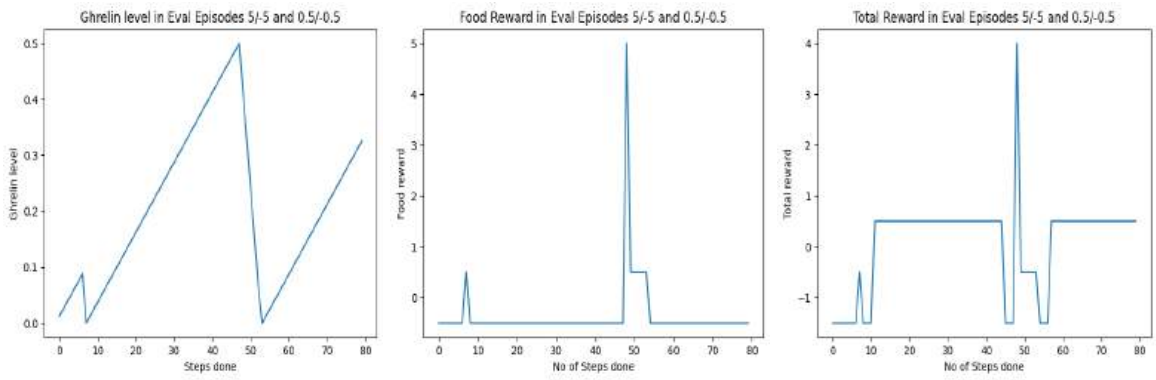



Figure 4.15: On the left, ghrelin levels of RL agent with heightened food reward of +5/-5, food located at (5,4) and hunger threshold is 0.5. In the center, food reward recieved by same RL agent. On the right, total reward received by the same RL agent.

### 4.2.3 Experiments investigating the influence of hunger thresholds

Two hunger thresholds are tested: 0.5 and 0.3 as detailed in Chapter 3 and the RL agent only experiences heightened food reward with it is "hungry". When hunger threshold is 0.5, the agent is able to take 40 steps (out of the permissible 80) before shifting from "satiated" to "hungry". In contrast, when hunger threshold is 0.3, the agent is only able to take 24 steps (out of the permissible 80) before it is "hungry" illustrated in Figure 3.4.

The behaviour of RL agents with a threshold of 0.5 is detailed in experiments under section 4.2.1. An RL agent with heightened food reward of +1/-1 with hunger threshold 0.3 behaves the same way as when its hunger threshold is 0.5 and food location (3,3). The agent never leaves the green goal square to revisit the food location even when its hunger threshold is exceeded because the total reward it receives (food reward + environment reward) is 0 as shown in Figure 4.6. This is higher than the reward for any available step outside the green goal square (-1).

The RL agent with heightened food reward of +2/-2, hunger threshold 0.3 and food located at (5,4) is able to manage its ghrelin levels by oscillating between the green goal square and the food location. From Chapter 3, every step within the food location is considered an eating instance or meal. The lower hunger threshold means the RL agent needs to take fewer meals (i.e., steps at food location) to become "fully satiated". Its trajectory is shown in Figure 4.18. The agent becomes "hungry" twice in a single episode (80 steps) as shown by its ghrelin levels in Figure 4.16.

As shown in Figure 4.17, the agent ventures out of the green goal square at approximately step 22 (2 steps before it reaches its hunger threshold of 0.3). It requires 3 steps to move from the green goal square to the food location. Thus, the agent becomes "hungry" before reaching the food location and it experiences a heightened food reward of -2 and +2, before and after reaching the food location respectively. The agent takes less steps at the food location (only 3) to reduce its ghrelin levels back to 0 because the threshold is lower as shown in Figure 4.16.

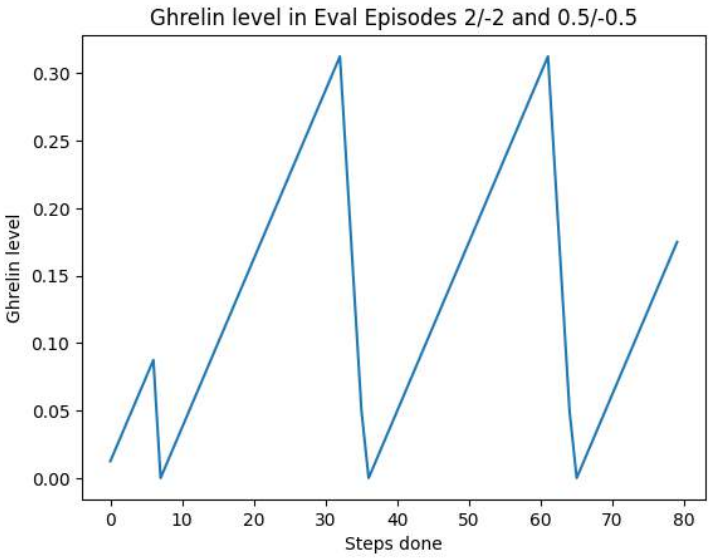



Figure 4.16: RL agent's ghrelin levels when hunger threshold is 0.3, heightened food reward +2/-2 and food located at (5,4).

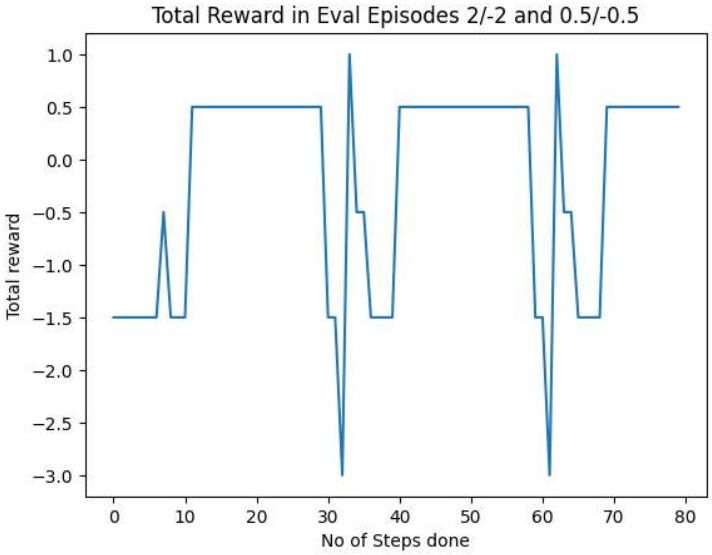



Figure 4.17: Total reward received by the RL agent with heightened food reward +2/-2, hunger threshold 0.3 and food located at (5,4).

When the same RL agent is deployed with heightened food reward +2/-2 and threshold 0.3 but food is located at (3,3), the agent remains at the food location enroute to the green goal square as shown in Figure 4.19. Its ghrelin levels remain 0 for the rest of the episode. The RL agent never reaches the hunger threshold and never receives heightened food reward. The agent endlessly rotates at the food location (gaining a reward of +0.5 for each step at the food location when "satiated") until the episode truncates. The static total reward illustrated in Figure 4.20 is calculated as follows:

$$-1_{env} + 0.5_{fr} = -0.5_{tr}$$

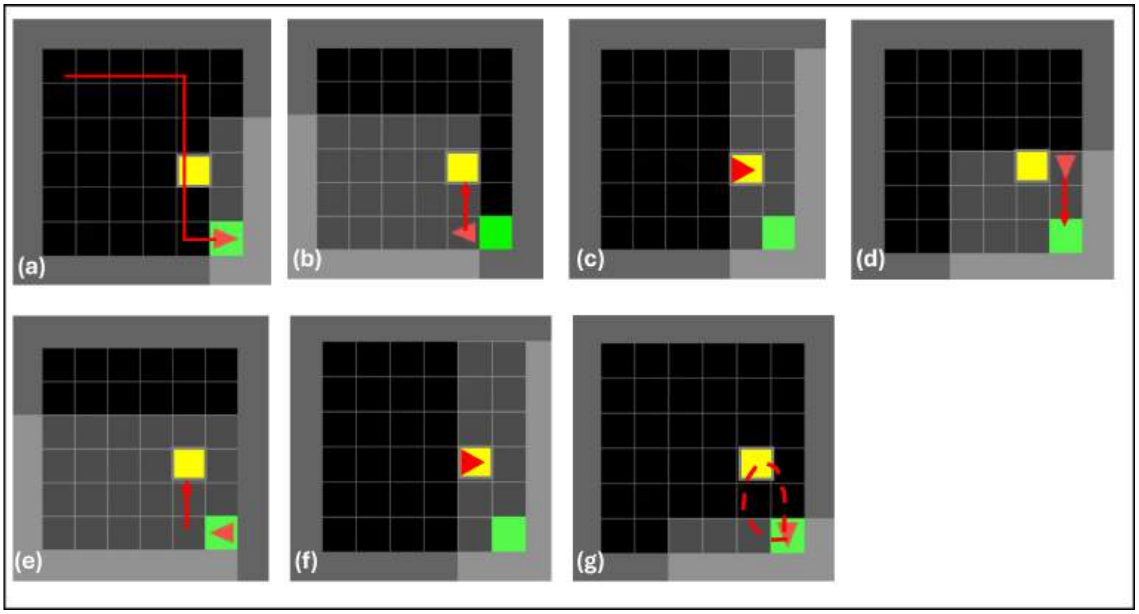



Figure 4.18: Learned policy of RL agent which receives a heightened food reward of +2/-2 above hunger threshold 0.3. Yellow square represents food located at position (5,4). From top left to bottom right, (a) Red line shows the trajectory of the agent at the start of the evaluation episode which passes food location (5,4). (b) After a few steps at the goal state (6,6), the agent leaves the green goal square in search of food as its ghrelin levels approach the hunger threshold 0.3.(c) The agent is officially "hungry" and enters the nearby food location (5,4) and takes approximately 3 steps in the food location which result in a decrease of ghrelin levels by 0.3, returning it to 0. The agent is now "fully satiated" again. (d) Afterwards, the agent returns to the green goal square goal. (e) After a few more steps at the green goal square, the agent nears its hunger threshold again and ventures back to the food location (5,4). It remains there for another 3 steps to decrease ghrelin levels back to 0. The agent is now "fully satiated" again. (g) The agent returns to the green goal square where it remains until the episode truncates.

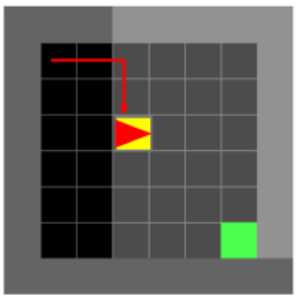


Figure 4.19: Learned policy by RL agent with heightened food reward +2/-2 and hunger threshold is 0.3. Red line illustrates trajectory of RL agent. Yellow square represents food location at (3,3)

An RL agent with heightened food reward +5/-5 and threshold 0.3 with food located at (3,3) behaves in the same way shown in Figure 4.19. The agent rotates endlessly

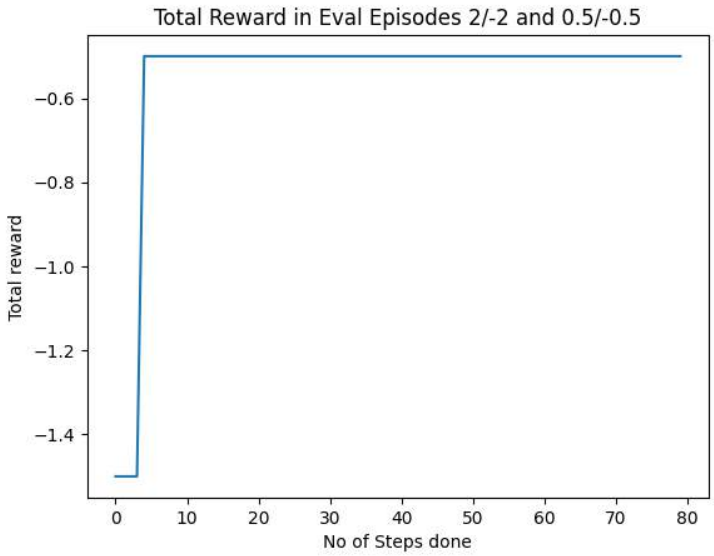



Figure 4.20: Total reward received by RL agent with heightened food reward +2/-2, food located at (3,3) and hunger threshold is 0.3.

at the food location enroute to the green goal square (receiving a food reward of +0.5 for each step) until the episode truncates. Its ghrelin level remains 0 for the rest of the episode. The RL agent never reaches the green goal square, its hunger threshold and never receives the heightened food reward. The total reward received by this agent is equivalent to the illustration in Figure 4.20. The difference in the RL agent's behaviour at hunger threshold 0.3 when the food location is closer to the goal square (5,4) and further away (3,3) suggests that food location and hunger threshold impact an agent's ability to manage its interoceptive needs (ghrelin levels). This requires additional experiments detailed in the next section 4.3.

## 4.3 Additional experiments

Ghrelin levels, food reward, hunger threshold and food location are four key factors to equip an RL agent with interoception. These key settings have initial values and functions. Adjusting these initial values and functions causes the agent to behave differently. As described in Chapter 3, the RL agent is deployed in the environment when it is "fully satiated" (i.e., ghrelin level is 0). Note that in every setting, the environment is embedded with a food location which is a specific point within the grid world for which ghrelin levels decrease by 0.1 (i.e., mimicking the effect of "eating food" on reducing hunger in living organisms). Lastly, note that the agent's ghrelin level is always between 0 and 1.

### 4.3.1 Proportional Food reward

To validate findings from experiments integrating baseline and heightened food reward as described in section 3.5.2, alternative experiments employing a food reward function

that is proportional to the agent's ghrelin levels are conducted. When in the food location, the RL agent receives a food reward that is proportional to its ghrelin level at the time (i.e., food reward equals positive ghrelin level) regardless of whether or not the RL agent is "hungry". Although this alternative set up still abides by a set hunger threshold (i.e., 0.5 or 0.3) it does not evoke a heightened food reward when "hungry". Instead, when outside the food location, the RL agent receives a food reward that is a negative value proportional to its ghrelin level shown in Figure 4.21 regardless of whether or not it is "hungry". In effect, food reward is always been -1 and 1, depending on the RL agent's ghrelin levels and if it is in the food location (i.e., "eating") or not.

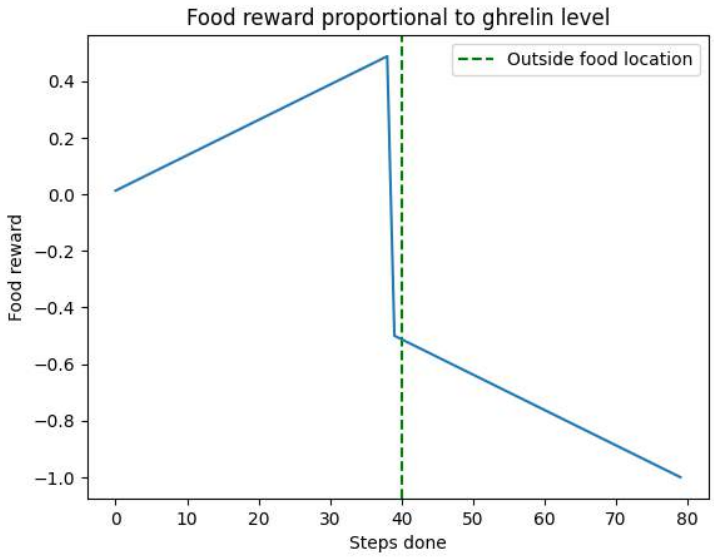



Figure 4.21: Alternative food reward function that is proportional to agent's ghrelin levels. Below the hunger threshold, food reward is equal to ghrelin level. Above hunger threshold, food reward is negative equivalent of ghrelin level.

This food reward function is applied to an RL agent deployed in a environment where food location is distant from the green goal square at position (3,3) and proximal to the goal state at position (5,4) as described in section 3.5.2 and shown in Figure 3.8 and Figure 3.10 respectively.

When an RL agent with a proportional food reward function is deployed in a grid-world environment with food located at position (3,3) with hunger threshold set to 0.5, the agent learns a policy which intersects the food location enroute to the green goal square. As shown in Figure 4.22, the agent remains at the green goal square until approximately step 76, where it ventures back to the food location causing the total reward to drop to approximately -1.9. This is because the episode truncates with the RL agent at position (3,6), enroute back to the food location (3,3). This is interesting behaviour because compared to the RL agent deployed with heightened food reward (+1/-1) with the same hunger threshold shown in Figure 4.6, the RL agent entirely ignores its ghrelin levels and never leaves the goal square despite becoming "extremely hungry" (i.e., ghrelin level equals 1). Even when the number of training episodes is

increased to 20,000, the RL agent still learns the same policy and ends the episode enroute to the food location at position (3,6).

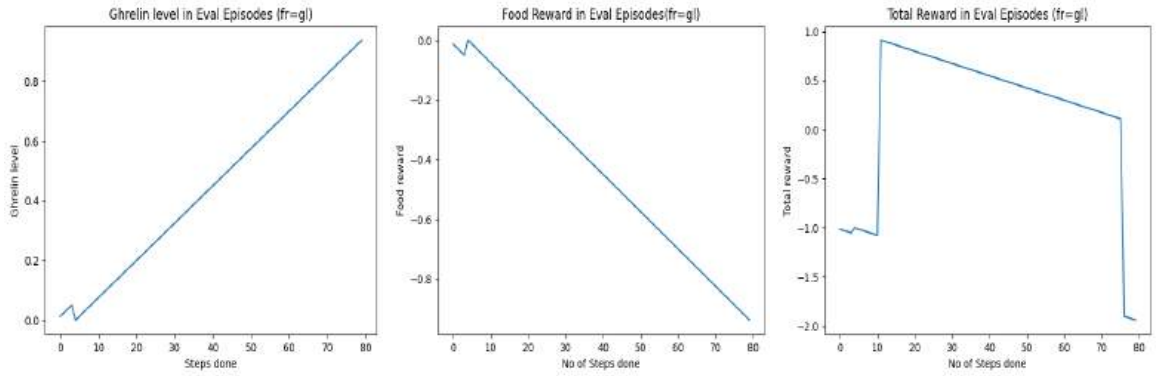



Figure 4.22: On the left, ghrelin levels of RL agent with proportional food reward, food located at (3,3) and hunger threshold is 0.5. In the center, food reward recieved by same RL agent during evaluation episode. On the right, total reward received by the same RL agent.

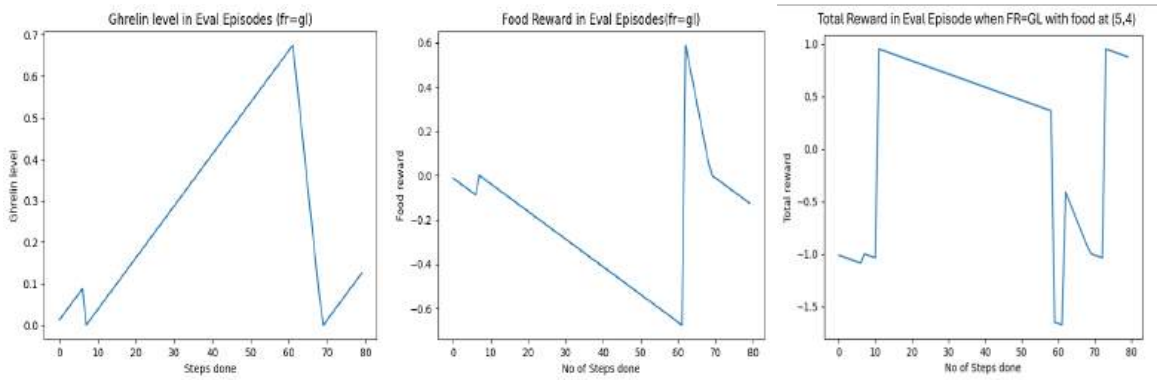



Figure 4.23: On the left, ghrelin levels of RL agent with proportional food reward, food located at (5,4) and hunger threshold is 0.5. In the centre, food reward received by same RL agent during evaluation episode. On the right, total reward received by the same RL agent.

An RL agent with a proportional food reward function shown in Figure 4.21 is deployed in an environment with a proximal food location at position (5,4) with a hunger threshold of 0.5. The RL agent learns a policy which intersects the food location (5,4) enroute to the green goal square. The RL agent remains at the green goal square as its ghrelin levels increase. This causes the total reward to steadily decrease from 1 as shown in Figure 4.23. This RL agent manages its ghrelin levels by visiting the food location after exceeding its hunger threshold as shown in Figure 4.23. When the RL agent's ghrelin levels approach 0.7, the agent leaves the green goal square to visit the food location. In the plot of total reward shown in Figure 4.23, this is the sharp switch from approximately -1.6 to -0.4, because food reward takes the positive value of ghrelin when the agent enters the food location as detailed below:

$$-1_{env} + 0.6_{fr} = -0.4_{tr}$$

It remains at the food location until it becomes ”fully satiated” (i.e., returning its ghrelin levels back to 0 which also brings food reward to 0 as well). Afterwards, the RL agent returns to the green goal square bringing the total reward to 1 where it remains until the episode ends. This behaviour is interesting but relatively expected given the ability of an RL agent with heightened food reward +1/-1 to manage its ghrelin levels when the food location was proximal to the goal state as shown in Figure 4.12.

An RL agent with a proportional food reward function shown in Figure 4.21 is deployed in an environment with food located at position (5,4) but with a hunger threshold of 0.3 as shown in Figure 3.4. The RL agent behaves similarly to when its hunger threshold is 0.5 as shown in Figure 4.23. It learns a policy which intersects the food location enroute to the green goal square. As shown in Figure 4.24, at approximately step 58, it leaves the goal state to ”eat” at the food location. After becoming ”fully satiated” (i.e., returning its ghrelin level to 0 and by extension food reward to 0), it returns to the green goal square. This is interesting behaviour because an RL agent with heightened food reward +1/-1 is unable to manage its ghrelin levels with the lower hunger threshold of 0.3 as detailed in section 4.2.3. Despite the reduced hunger threshold which should cause the RL agent to ”eat” more frequently, this RL agent remains at the green goal square until approximately the same time step as when its hunger threshold was 0.5. This means the RL agent is not adequately managing its ghrelin levels because its not eating as frequently as it should with a hunger threshold of 0.3. This suggests a decreased influence of hunger threshold on an RL agent's ability to manage its ghrelin levels when the food reward function is proportional compared to when food reward has two values (i.e., baseline and heightened food reward values).

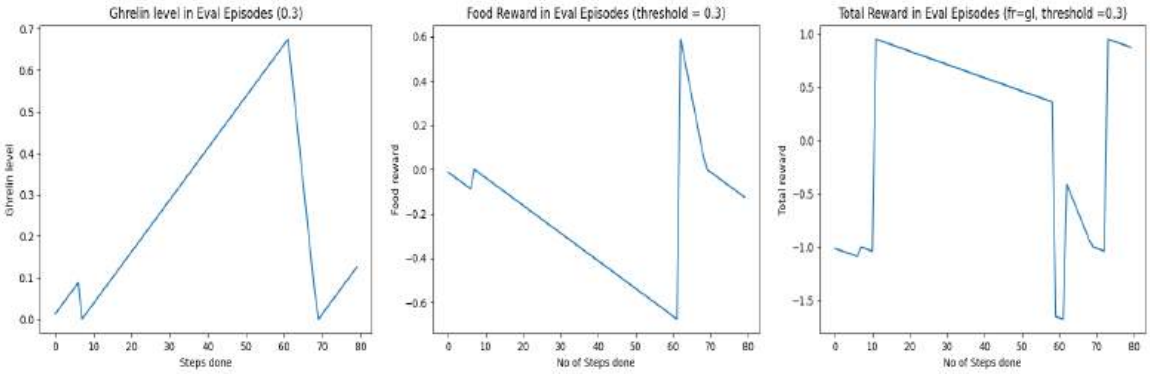



Figure 4.24: On the left, ghrelin levels of RL agent with proportional food reward, food located at (5,4) and hunger threshold is 0.3. In the centre, food reward received by same RL agent during evaluation episode. On the right, total reward received by the same RL agent.

### 4.3.2 RL agent with Ghrelin Control Actions

This experiment equips an RL agent with additional actions to internally control its ghrelin levels without needing to visit a food location as detailed in Chapter 3. The RL agent receives information of its internal ghrelin levels, but also possess two additional actions:

- 3 - reduce ghrelin level by 0.1
- 4 - increase ghrelin level by 0.1

These additional actions required a slight adjustment of the DQN architecture shown in Figure 3.6 to include two additional output nodes. These additional actions do not alter the orientation or position of the RL agent and only affect its internal ghrelin levels. Ghrelin levels are still capped between 0 and 1, even if the agent continues to select action 3 or 4 at these limits. To investigate if the RL agent would visit the embedded food location or utilise its internal actions to control its ghrelin levels, experiments with ghrelin control actions also integrate a food location, baseline and heightened food reward. The RL agent only receives the heightened food reward for being in the embedded food location as detailed in section 4.2.2.

An RL agent with ghrelin control actions is deployed an environment with food located at position (3,3) on the grid. The RL agent receives a baseline food reward of +0.5/-0.5 and heightened food reward of +2/-2. The hunger threshold is set to 0.5. The RL agent learns a policy that does reaches the green goal square without intersecting the food location (3,3) - the same trajectory as shown in Figure 4.3. As shown in Figure 4.25, the RL agent keeps its ghrelin levels below the hunger threshold by utilising action 3 (decrease ghrelin) after reaching the green goal square. The food reward is 0.5 as a consequence of capping ghrelin levels from being negative. When the agent "eats", its ghrelin decreases by 0.1, but when its ghrelin level is already below 0.1 and the agent eats, it gets a food reward of 0.5. This is shared in previous experiments with food location (when eating was only possible in the food location as described in section 4.2.2), and was applied in this iteration as well. The total reward is:

$$1_{env} + 0.5_{fr} = 1.5_{tr}$$

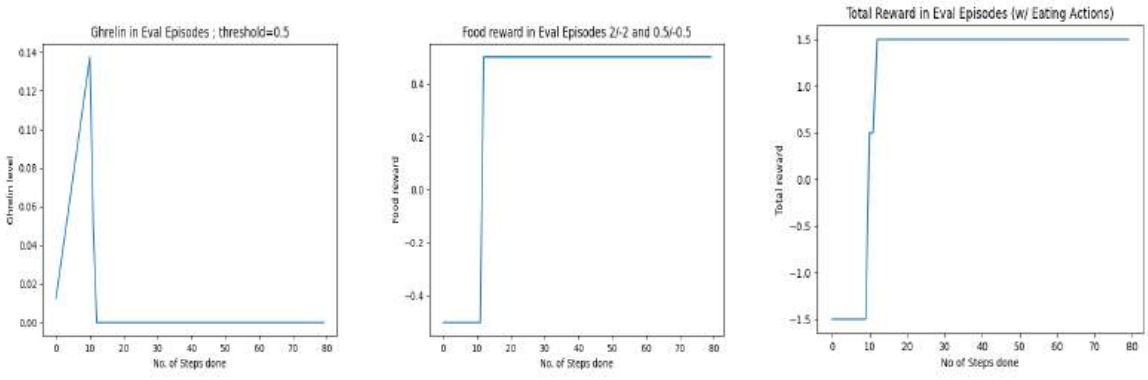



Figure 4.25: On the left, ghrelin levels of RL agent with ghrelin control actions, food located at (3,3) and hunger threshold is 0.5. In the centre, food reward received by same RL agent during evaluation episode. On the right, total reward received by the same RL agent.

An RL agent with ghrelin control actions is deployed with a proportional food reward function described in section 4.3.1 to further investigate the influence of heightened

food reward on the agent's learned policy. In this experiment, food is located at position (3,3) and hunger threshold is set to 0.5. The RL agent learns the same optimal policy as shown in Figure 4.3 despite the presence of a food location. The RL agent reaches the green goal square and manages its ghrelin levels by selecting action 3 (decrease ghrelin level) until the episode ends. This is shown in Figure 4.36. The total reward is 1, because the food reward is proportional to ghrelin levels which is constantly 0, so the RL agent only receives the environment reward for being at the green goal square.

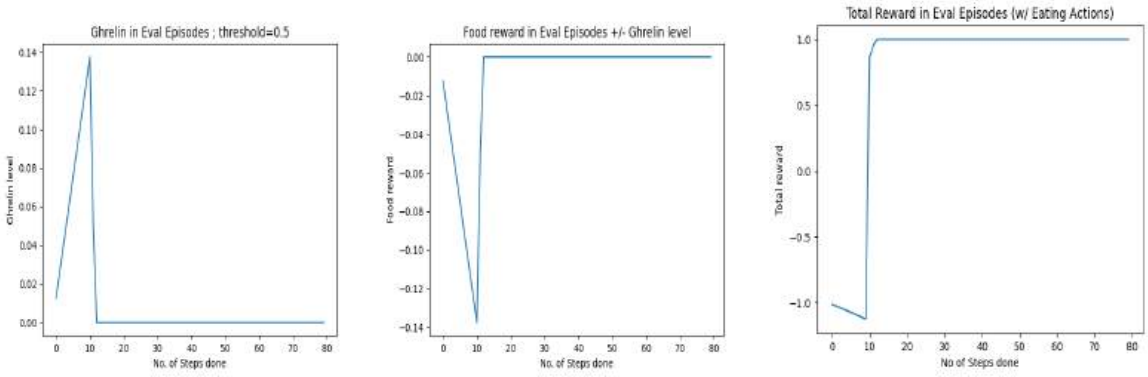



Figure 4.26: On the left, ghrelin levels of RL agent with ghrelin control actions, food located at (3,3) and hunger threshold is 0.5. In the centre, food reward received by same RL agent during evaluation episode which is proportional to its ghrelin levels as described in section 4.3.1. On the right, total reward received by the same RL agent.

These results suggest that when an RL agent has internal actions to control its ghrelin levels, food location has no influence on its learned policy. This is demonstrated by the RL agent taking the baseline path to the green goal square unlike the routes detailed in section 4.2.2. It is also interesting that the RL agent does not utilise its ghrelin control actions to exploit the environment for the heightened food reward, attainable at the food location beyond its hunger threshold.

### 4.3.3 Starting episode with RL agent "hungry"

When the RL agent begins the episode "hungry" (i.e., ghrelin level equal to hunger threshold), the agent takes additional steps at the food location enroute to the green goal square before proceeding. These additional steps involve rotating (i.e., oscillating left and right) at the food location which each cause a decrease of 0.1 in ghrelin levels. In essence, the agent is "eating" until its ghrelin levels return to 0 before continuing to the green goal square. This behaviour is the same across all three heightened food reward values. This behaviour in response to hunger thresholds is unique for the RL agent that receives a heightened food reward of +1/-1 because when this agent is deployed "fully satiated" (ghrelin level = 0) as in section 4.2.1, the agent never revisits the food location (3,3) after reaching the green goal square despite exceeding its hunger threshold.

### 4.3.4 Removing effect of food

When the defined food location does not cause a decrease in the agent's ghrelin level, the agent remains at the food location until the episode truncates and never reaches the green goal square. This behaviour is the same for all heightened food reward values, food locations and thresholds. This means the agent is unable to attain the extrinsic reward of the green goal square if it is "hungry". An example of ghrelin, food reward and total reward experienced by an RL agent which receives heightened food reward of +2/-2 is shown in Figure 4.27 in this set up. Though the food location does not cause a decrease in ghrelin levels, the agent still receives a heightened food reward for being in the food location when hungry. As shown in Figure 4.27, the RL agent remains at the food location until its hunger threshold is exceeded and it begins to receive the heightened food reward of +2 for being at the food location. Although this conveys an overriding influence of interoception on the RL agent's behaviour, the underlying food reward dynamics are causing this behaviour. While food location does affect the agent's learned policy as demonstrated by experiments detailed in section 4.2.2 in terms of the path it takes to reach the green goal square, this experiment demonstrates that toggling between heightened and baseline food reward values is necessary to recreate the shifting hormones and value of food when hungry versus satiated as detailed in section 2.5 which ensures that the RL agent does not fixate on state that maximises reward.

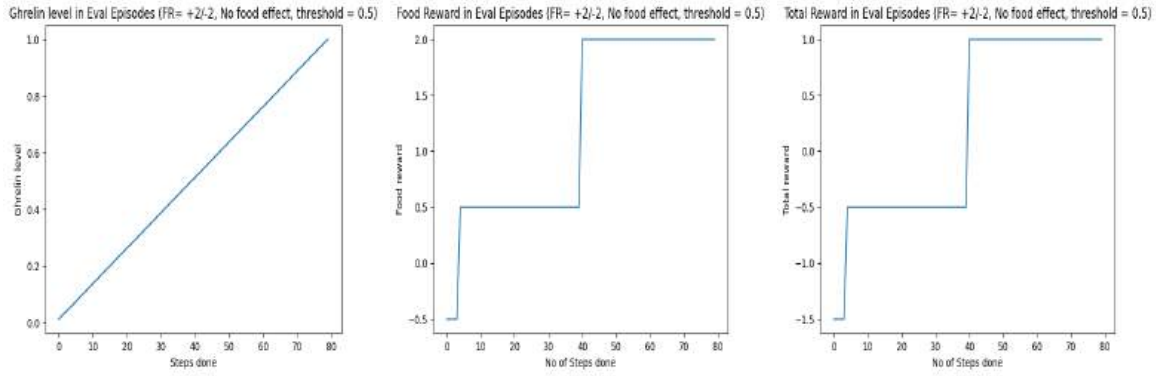



Figure 4.27: On the left, ghrelin levels of RL agent which receives heightened food reward +2/-2, food located at (3,3) but does not cause a decrease of ghrelin and hunger threshold is 0.5. In the centre, food reward received by same RL agent during evaluation episode. On the right, total reward received by the same RL agent.

### 4.3.5 Extreme ghrelin levels

When the RL agent begins the episode extremely "hungry", with its ghrelin level being double the set hunger threshold, the RL agent's ability to reach the food location or the green goal square depends on its heightened food reward. For an RL agent that receives a heightened food reward of +1/-1, trained on a hunger threshold of 0.5, and food located at position (3,3), deployed at ghrelin levels double this threshold (i.e., 1), it remains at the starting position (1,1).

For an RL agent that receives a heightened food reward of +2/-2, trained on a hunger threshold of 0.5, and food located at position (3,3), deployed at ghrelin levels double this threshold (i.e., 1), the agent is able to take additional steps (i.e., oscillating left and right to ”eat”) at the food location enroute to the green goal square. Each step within the food location causes a decrease of ghrelin levels by 0.1 and this enables the agent to become ”satiated” before continuing to the green goal square. Its subsequent behaviour is the same as illustrated in Figure 4.5. The agent revisits the food location when its hunger threshold is exceeded and returns to the green goal square when it is ”satiated”. The same behaviour is exhibited when the RL agent receives a heightened food reward of +5/-5.

### 4.3.6 Arbitration of intrinsic versus extrinsic reward

The behaviour of the agent when it begins the episode “hungry” is interesting for a number of reasons. It confirms that the intersection of the food location enroute to the green goal square is intentional by the RL agent as further supported by results in section 4.3.2. This is because when it starts “hungry”, the RL agent takes additional steps (”eating” at the food location) to reset its ghrelin levels before proceeding to the green goal square as detailed in section 4.3.3. This is also the case when heightened food reward is comparable to the environment’s reward (i.e., +1/-1) and the agent under normal conditions (starting “fully satiated”) ignores its interoceptive needs later on (detailed in section 4.2.1). To investigate the arbitration of intrinsic interoceptive reward associated with the food location and extrinsic reward associated with the green goal square, an RL agent with hunger threshold 0.5, heightened food reward +2/-2 is deployed in an environment where food location and the green goal square are equidistant as illustrated in Figure 4.28.

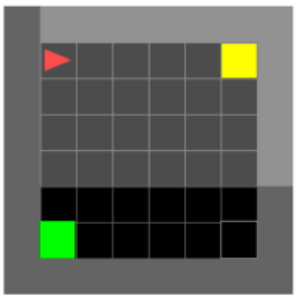


Figure 4.28: View of grid world in experiment where food location (yellow square) and green goal square are equidistant. RL agent receives heightened food reward of +2/-2 and hunger threshold is 0.5.

When the agent is deployed “fully satiated” (ghrelin level = 0 at start of episode), it visits the green goal square before venturing to the food location when it becomes “hungry”. After 5 steps in the food location, reducing ghrelin levels back to 0 (i.e. becoming “fully satiated”), the agent returns to the green goal square. It remains there until the episode truncates. When this same RL agent is deployed “hungry”, the opposite behaviour is shown. The agent first visits the food location, remains there until it is

"fully satiated" (i.e. its ghrelin levels return to 0) before proceeding to the green goal square. This behaviour is facilitated by the toggling of heightened and baseline food reward values when the RL agent is hungry versus satiated inspired by living organisms detailed in section 2.5.

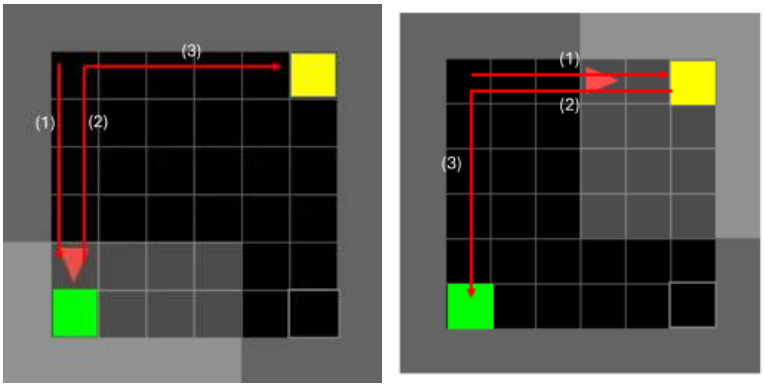



Figure 4.29: On the left, trajectory of "fully satiated" RL agent when food location (yellow square) and green goal square are equidistant. Hunger threshold is 0.5. Numbered arrows are the sequential trajectory of RL agent. On the right, trajectory of the same RL agent "hungry".

The RL agent's behaviour is consistent with the survival-based behaviour of a real-life hungry mouse because it prioritises reaching the food location (i.e., mouse eating the block of cheese) before the green goal square (i.e., before running on its wheel). In earlier experiments, a lower hunger threshold evidently affects the learned behaviour of the RL agent by causing it to oscillate between the food location and green goal square. It is important to investigate if this survival-based behaviour is also mimicked by the RL agent when its hunger threshold is decreased from 0.5 to 0.3.

The same RL agent which receives a heightened food reward of +2/-2 is deployed to the adjusted environment above where food location and green goal square are equidistant (Figure 4.28) except with a hunger threshold of 0.3. The agent moves to the food location at (6,1) where it remains until the episode truncates. The RL agent never ventures to the green goal square. Its trajectory is shown in Figure 4.30.

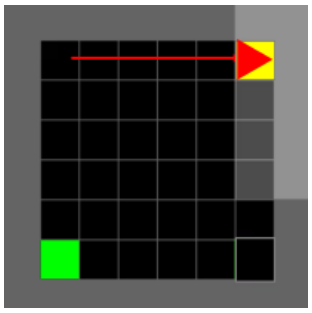


Figure 4.30: Red line is trajectory taken by RL agent when food location (yellow square) and green goal state are equidistant and hunger threshold 0.3.

In earlier experiments detailed in section 4.2.2, the position of the food location in the grid world affected the RL agent's ability to effectively manage its ghrelin levels by visiting the food location before or shortly after becoming "hungry". It is important to investigate if the agent above is unable to manage its ghrelin levels because food location is a significant distance from the green goal state.

The same RL agent is deployed "fully satiated" to the adjusted environment where the green goal square is located at position (1,6) and food location is adjusted to a position closer to the green goal state at (2,4) shown in Figure 4.31). The agent receives a heightened food reward of +2/-2 and its hunger threshold is 0.3. The RL agent visits the green goal square and then visits the food location when it becomes "hungry" (i.e. when its ghrelin levels exceed the hunger threshold). Its full trajectory is illustrated in Figure 4.32. In effect, the agent is able to manage its interoceptive needs (ghrelin levels) while also reaching the green goal square. The episode ends (truncates) when the agent is enroute to revisit the food location for a third time.

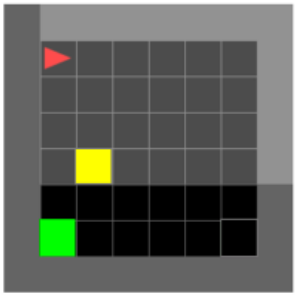


Figure 4.31: View of adjusted grid world where green goal square is located at (1,6) and food location (yellow square) is (2,4)

When this RL agent is deployed "hungry", it first visits the food location and then moves to the green goal square. It oscillates between the two depending on its ghrelin level approaching the set hunger threshold (0.3). The episode ends (truncates) with the RL agent at the green goal square. The agent's full trajectory is illustrated in Figure 4.33. This demonstrates that the agent can better balance intrinsic (interoceptive needs by visiting the food location) and extrinsic (reward of reaching the green goal square) reward when these reward sources are proximal.

### Arbitration when food reward is proportional to ghrelin levels

An RL agent which possess a proportional food reward function as shown in Figure 4.21 is deployed in the equidistant gridworld shown in Figure 4.28. The hunger threshold is set to 0.5. This RL agent learns a policy that reaches the green goal state at (1,6) but never visits the food location at (6,1) despite exceeding its hunger threshold as shown in Figure 4.34. The RL agent receives a static total reward of 0 when the episode truncates. Even when this RL agent is deployed "hungry" (i.e., ghrelin level equals hunger threshold 0.5), its behaves in the same way. The agent reaches the green goal square where it remains until the end of the episode despite exceeding its hunger threshold.

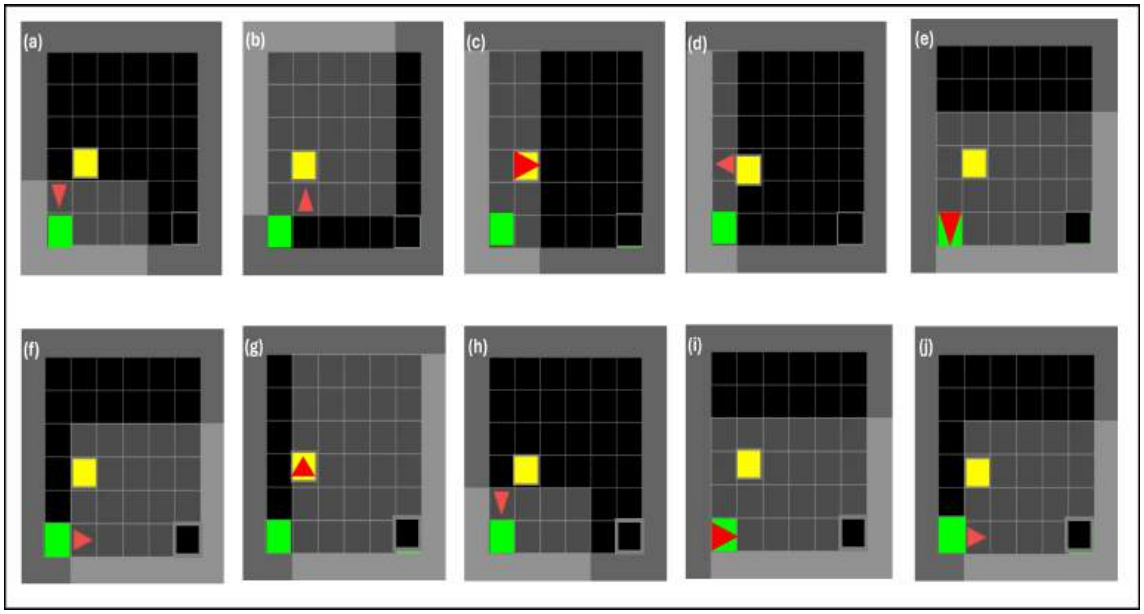



Figure 4.32: Frames of learned policy by "satiated" RL agent which receives a heightened food reward of +2/-2 above hunger threshold 0.3. Yellow square represents food located at position (2,4) which is closer to the adjusted green goal square at position (1,6).

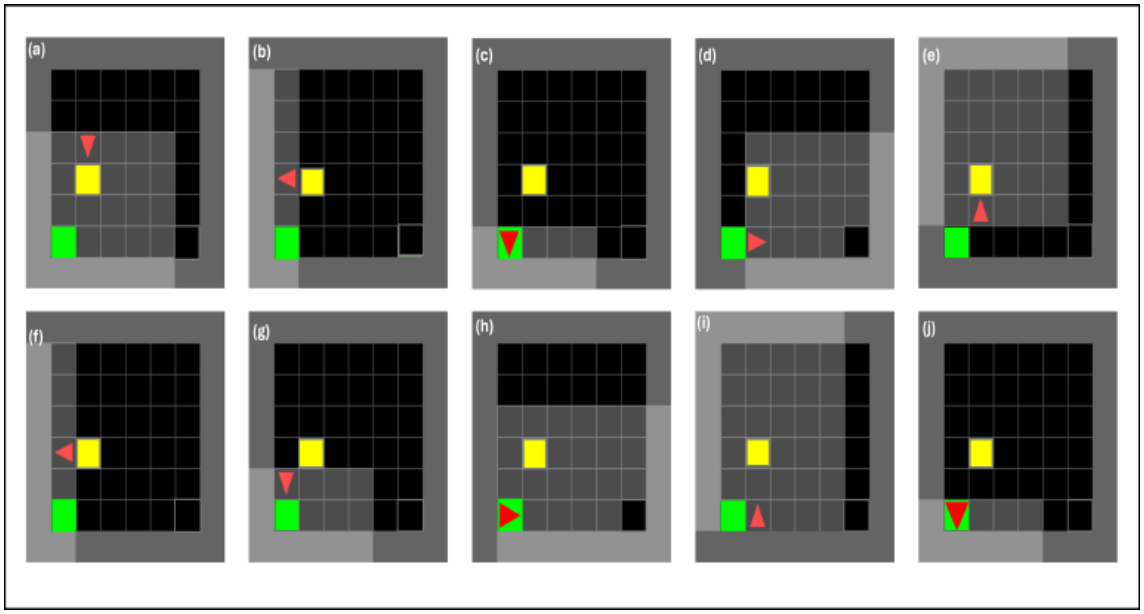



Figure 4.33: Frames of learned policy by "hungry" RL agent which receives a heightened food reward of +2/-2 when deployed at hunger threshold 0.3. Yellow square represents food located at position (2,4) which is closer to the adjusted green goal square at position (1,6).

### Arbitration when agent has ghrelin control actions

An RL agent with actions to internally control its ghrelin levels as described in section 4.3.2 is deployed in an equidistant gridworld as shown in Figure 4.28. The food loca-

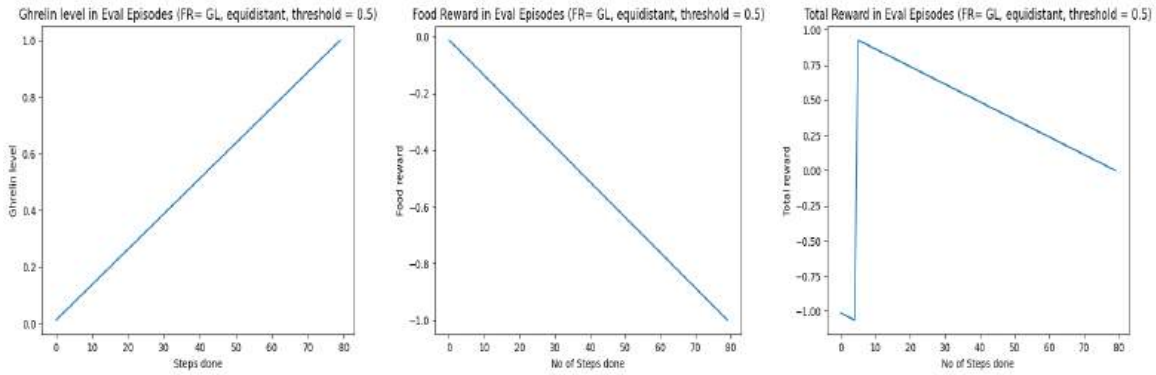



Figure 4.34: On the left, ghrelin levels of RL agent with a proportional food reward function deployed in an equidistant gridworld. In the centre, food reward received by same RL agent during evaluation episode. On the right, total reward received by the same RL agent.

tion is adjusted to position (6,1) in the gridworld and hunger threshold is set to 0.5. The RL agent receives a baseline and heightened food reward of +0.5/-0.5 and +2/-2 respectively for being at the food location when satiated versus hungry. This RL agent remains at the starting position (1,1) where it continues to select action 4 (increase ghrelin level) until the episode truncates as shown in Figure 4.35. It does not reach the green goal square nor the food location and continued to receive a total reward of -3 until the episode truncates. Even when this agent is deployed ”hungry” (i.e., ghrelin levels equals hunger threshold of 0.5), the agent behaves the same way.

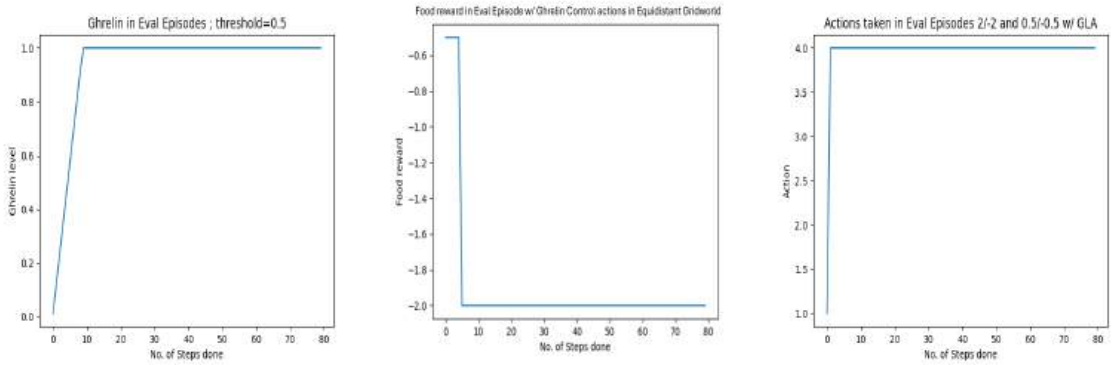



Figure 4.35: On the left, ghrelin levels of RL agent with ghrelin control actions deployed in an equidistant gridworld. In the centre, food reward received by same RL agent during evaluation episode. On the right, actions taken by RL agent during evaluation episode.

The behaviour of an RL agent with proportional food reward and ghrelin control actions in an equidistant gridworld respectively are interesting. In both scenarios described above, the RL agent is unable to manage its ghrelin levels internally nor visit the food location despite exceeding its hunger threshold in both settings. This is poor management of ghrelin levels compared to the performance of these agents in a normal

gridworld environment shown in Figure 3.1. To examine the combined performance of both settings (i.e., having internal ghrelin control actions and a proportional food reward function) in an equidistant gridworld, additional experiments are conducted.

An RL agent with actions to internally control its ghrelin levels described in section 4.3.2 and a proportional food reward function described in section 4.3.1 is deployed in an equidistant gridworld shown in Figure 4.28. The hunger threshold is set to 0.5, food is located at position (6,1) and the green goal square is at position (1,6) in the gridworld. This RL agent learns a policy to reach the green goal square first and then manages its ghrelin levels using its internal action 3 (decrease ghrelin level). When this RL agent is deployed "hungry" (i.e., ghrelin level equals hunger threshold 0.5), it behaves the same way. It first reaches the green goal square and then begins to take action 3 to control its ghrelin levels as shown in Figure 4.36.

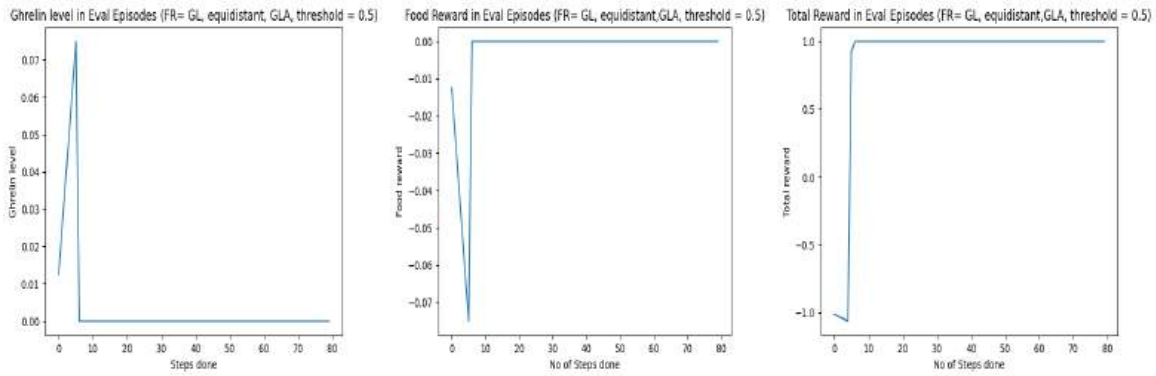



Figure 4.36: On the left, ghrelin levels of RL agent with ghrelin control actions deployed in an equidistant gridworld with a proportional food reward function. The agent begins the episode "fully satiated" (i.e., ghrelin level is 0). In the centre, food reward received by same RL agent during evaluation episode. On the right, total reward received by this RL agent during evaluation episode.

When this RL agent is deployed "extremely hungry" (i.e., ghrelin level is double threshold, 1), it first takes action 3 to decrease its ghrelin level to approximately 0.5 (i.e., the set hunger threshold), then it continues to the green goal square where it remains until the episode truncates while selecting action 3 (which keeps ghrelin levels at constant 0 and by extension food reward to 0) as shown in Figure 4.37. This is interesting behaviour because an RL agent with only ghrelin control actions in an equidistant gridworld fails entirely to control its ghrelin levels as shown in Figure 4.35. Similar to experiments with internal ghrelin control actions described in section 4.3.2, the RL agent learns a policy that never visits the food location. This confirms that having internal ghrelin control actions eliminates the need for a pre-defined food location. It also illustrates that with ghrelin control actions, an RL agent is only able to prioritise intrinsic interoceptive needs over extrinsic rewards in extreme settings.

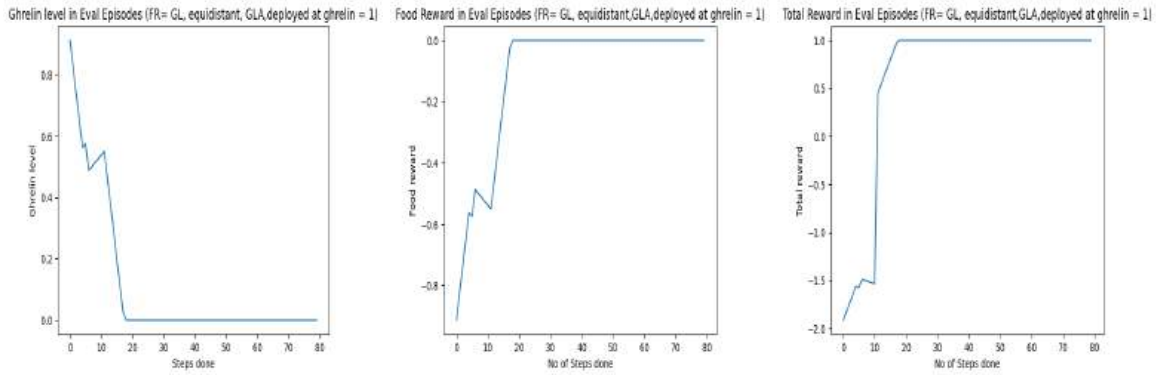



Figure 4.37: On the left, ghrelin levels of RL agent with ghrelin control actions deployed in an equidistant gridworld with a proportional food reward function. The agent begins the episode ”extremely hungry” (i.e., ghrelin level is 1). In the centre, food reward received by same RL agent during evaluation episode. On the right, total reward received by this RL agent during evaluation episode.

### 4.3.7 Removing food reward

The above experiments are repeated removing the food reward the agent receives for visiting the food location (both heightened and baseline food reward) while maintaining its effect on ghrelin levels. The RL agent learns the same optimal policy detailed in sections 4.1.1 and 4.1.2 when interoceptive signals (i.e. ghrelin levels) are omitted. The RL agent moves directly from the starting position to the green goal square as illustrated in Figure 4.3.

### 4.3.8 Increased function of food location

In this study, the effect of food lowering ghrelin levels in the human body was modeled as a decrease of ghrelin levels by 0.1 (see Chapter 3) for every eating instance or ”meal”. This effect is fixed across all experiments. To investigate if the RL agent can learn a preference for food type (i.e. foods which lower ghrelin levels to different extents), an RL agent with heightened food reward +2/-2 and hunger threshold 0.5 is deployed in an environment with two food locations which have different effects on ghrelin levels. The first location is (2,5) (represented by a yellow square in Figure 4.3) and causes the standard decrease of ghrelin levels by 0.1. The second location is (5,2) (represented by a purple square in Figure 4.3) and causes a heightened decrease of ghrelin levels by 0.3. This adjusted environment is shown in Figure 4.3. The agent learns a policy which intersects the second food location as shown in Figure 4.39 and behaves like the agent in section 4.2.1 with the same settings. Meaning that the agent intersects the second food location enroute to the green goal square and revisits this second food location when its hunger threshold is exceeded and returns to green goal square when it is ”fully satiated” again.

In conclusion, varying the effect of the food location on the agent’s ghrelin levels can be used to model the biological responses and preferences of different food types which affect ghrelin levels as shown by Erdmann *et al.* [2003].

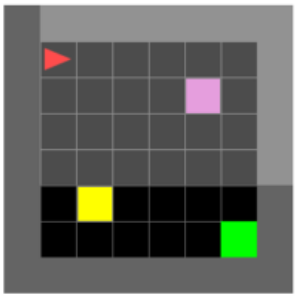


Figure 4.38: View of adjusted grid world at start of episode. Red triangle is RL agent, green square is goal green square. Standard food location is represented as yellow square at position (2,5) and second food location is represented as purple square at (5,2).

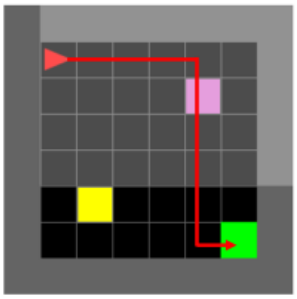


Figure 4.39: Red line shows first half of learned policy by RL agent deployed in environment with two food locations with different effects on ghrelin levels. Yellow square is standard food location which causes decrease of 0.1. Purple square is second food location which causes decrease of ghrelin by 0.3.

# Chapter 5

# Discussion

Interoception plays a key role in the survival of a living organism [Tsakiris and Critchley 2016]. To maintain homeostasis, an organism responds to interoceptive signals conveying its basic needs like eating, drinking, and sleeping. Interoceptive needs like eating require an organism to take external actions in its environment which requires the formation of an action plan to meet these needs [Liu and Kanoski 2018]. The brain regulates meeting interoceptive needs without entirely compromising other activities by a holistic evaluation process [Averbeck and Murray 2020; Kim 2013]. This is done by an interplay of the mesolimbic and mesocortical systems in the brain. The prefrontal cortex (PFC), and specifically the orbitofrontal cortex (OFC), amygdala and nucleus accumbens (NAcc) subsystem, processes interoceptive signals received from the hypothalamus (the homeostatic centre of the body) and evaluate possible future actions depending on these signals [Ferrario *et al.* 2016; Kim 2013].

Modelling this evaluation process done at the PFC requires the complete integration of interoceptive signals. To do this, the RL agent and its environment are equipped with four key characteristics: (1) ghrelin levels, (2) hunger threshold, (3) food location and (4) food reward. The agent receives a ghrelin level signal which reflects its own ghrelin levels modeled as a linearly increasing variable from 0 to 1 over the course of a single episode (80 steps). The agent possesses a fixed hunger threshold. This is the value of ghrelin level which switches the agent from “satiated” (ghrelin levels below hunger threshold) to “hungry” (ghrelin levels above hunger threshold). This study investigates two hunger thresholds: 0.5 and 0.3 which allow the RL agent 40 and 24 steps within the environment respectively before it switches to “hungry” (shown in Figure 3.4). To model the effect of food on lowering ghrelin levels, a pre-defined location (state) within the grid world is set as food location. The agent’s ghrelin levels decrease by a fixed 0.1 for every step (action) within this food location. Every step at the food location is considered an eating instance or ”meal”. To reflect the comprehensive evaluation process done by the OFC-Amygdala-NAcc subsystem, the RL agent receives two distinct values of food reward when it is “satiated” versus “hungry”. When the agent is “satiated”, it receives a baseline food reward of +0.5 and -0.5 for being at the food location and vice versa. This value is less than the base environment’s reward for the agent’s actions with respect to getting to the green goal square (+1/-1). This is purposely selected to reflect

the sub-par rewarding value of fodo when an organism is satiated compared to externally rewarding tasks based on the literature detailed in section 2.5. When the agent is "hungry", it receives a heightened food reward. To investigate this comprehensive valuation process of food when the agent is "hungry", this study experiments with three main values of heightened food reward: +1 and -1; +2 and -2; +5 and -5. Positive values are what the agent receives for being at the food location when "hungry" versus being in any other location (or state) within the grid (including the green goal square) and vice versa.

The obtained results show that the holistic evaluation of food (encompassed in the food location) as more rewarding than any other state within the environment is necessary to accurately model a healthy mouse's behaviour. This is shown by the behaviour of an RL agent with heightened food reward of +2/-2 leaving the green goal square to visit the food location just before its hunger threshold is exceeded and returning once its ghrelin levels return to 0 (shown in Figure 4.19). In extreme cases of heightened food reward of +5/-5, the RL agent learns a policy to pre-emptively visit the food location before exceeding its hunger threshold by taking into account the number of steps required to move from the green goal square to the food location (shown in Figure 4.8). This agent is also able to return to the green goal square once its ghrelin levels are returned to 0 after taking steps (i.e. "eating") at the food location. When agents with heightened food reward of +2/-2 and +5/-5 are given a lower hunger threshold (0.3), they exhibit the same ability to manage their interoceptive signals and needs (i.e. ghrelin levels and hunger thresholds) by oscillating between the pre-defined food location and green goal square more frequently when this food location is closer to the green goal square. With a proximal food location and moderate hunger threshold, RL agents which receive a heightened food reward of +1/-1 and proportional food reward function are also able to effectively manage their ghrelin levels by visiting the food location.

A relationship between food location and hunger threshold is demonstrated by the results obtained. When hunger thresholds are low (i.e., hunger threshold 0.3) and food location is a significant distance from the green goal square (i.e., food located at (3,3)), the RL agent no longer behaves as a healthy mouse. In these experiments, the RL agent shows signs of obesity by remaining at the food location despite always being "fully satiated" (ghrelin level = 0) [Stover *et al.* 2023; Ferrario *et al.* 2016]. Obesity has been extensively studied for the dysregulation of reward processing associated with food and consummatory behaviour in the absence of relevant hunger signals (i.e. ghrelin release in the stomach) [Stover *et al.* 2023; Baik 2021; Murray *et al.* 2014].

Other clinical comparisons appear in RL agents which receive a heightened food reward of +1/1 which is equivalent to the base environment's reward (+1 for reaching the green goal square and -1 for any other step). RL agents with heightened food reward of +1/-1 do not address their interoceptive needs despite receiving a negative food reward. This agent remains at the green goal square regardless of the hunger threshold being exceeded (tested with 0.5 and 0.3). This is consistent with feeding patterns of depressed patients with anhedonia [Murray *et al.* 2021]. Anhedonia reduces

the ability of a patient to experience pleasure [Coccurello 2019]. This reduced pleasure is reflected in the RL agent's heightened food reward value when it is "hungry" being equivalent to other available states (i.e. the base environment's +1 and -1 for being at the green goal square). In this analogy, the green goal square can represent any activity that does not address interoceptive needs which may be lying in bed rather than a strenuous activity like running on a wheel for a mouse.

The obtained results also show that interoceptive signals influence the learned policy of the RL agent regardless of the extremity of heightened food reward (+1/1; +2/-2;+5/-5). This is demonstrated by RL agents consistently learning a policy which intersects the food location enroute to the green goal square regardless of where it is placed in the grid world. This is further validated by RL agents equipped with internal ghrelin control actions learning a policy which never visits the food location. When these agents are deployed at a deficit (i.e. starting the episode "hungry" rather than "satiated") at a reasonable hunger threshold (i.e., 0.5), they are able to flexibly take additional steps at the food location (i.e., "eating" first) to become "satiated" before proceeding to the green goal square. In these experiments, toggling between heightened and baseline food reward dependent on an agent's ghrelin levels was crucial to recreating healthy eating behaviours observed in mice.

To answer the main question of this study, the PFC's arbitration of intrinsic versus extrinsic reward, results obtained in experiments where food location and the green goal square are equidistant show that actions of the RL agent are dependent on its internal state. From Chapter 4, when the RL agent is deployed "fully satiated" in this equidistant variant of the environment (shown in Figure 4.28), it visits the extrinsically rewarding green goal square and then the intrinsically rewarding food location when it becomes "hungry". In contrast, when this RL agent is deployed "hungry" in the same environment, it first visits the intrinsically rewarding food location and after becoming "satiated", it proceeds to the green goal square. This mimics the high-level valuation process done at the OFC-Amygdala-NAcc subsystem (which this study models as a single unit) when an organism is hungry by prioritising possible actions depending on interoceptive signals (and by extension, the organism's survival needs). It then prioritises an organism's actions primarily on which actions respond to this survival-based intrinsic reward and then based on which actions are externally rewarding. This parallel is facilitated by associating baseline and heightened food reward values when an RL agent is "hungry" versus "satiated" respectively. Additional experiments demonstrate that an RL agent with internal ghrelin control actions is only able to recreate this natural prioritisation of intrinsic needs under "extreme hunger" in an environment where intrinsic and extrinsic reward sources are equidistant.

## 5.1 Limitations

The RL agents trained above behave similar to the naturally observed behaviour of mice in neuroscientific literature but this relies on the following:

1. The environment during evaluation must match the training environment. This is shown by the behaviour of the RL agent when food location does not cause a decrease in ghrelin levels (i.e. agent gets stuck at food location until episode truncates).

2. The hunger threshold which the agent trains on must match the hunger threshold during evaluation. Even when this value is altered during evaluation, the RL agent's learned policy is not affected and the agent will visit the food location whether it is "hungry" or not depending on its ghrelin level and pre-defined hunger threshold during training.

3. The RL agent will continue to visit the food location to keep its ghrelin levels below the hunger threshold even if it does not receive a food reward during evaluation. This can give the misconception that food reward is not a contributing factor in the RL agent's learned ability to manage its ghrelin levels.

## 5.2 Future work

This study models ghrelin levels as a linearly increasing variable over time and hunger thresholds as a fixed point. The results obtained in Chapter 4 demonstrate that RL agents always remain at the food location until they are "fully satiated" by returning ghrelin levels back to 0. Though, any ghrelin level before the hunger threshold theoretically shifts the agent back to "satiated". This could be because of the immediate switch between heightened food reward and baseline food reward which models the OFC-Amygdala-NAcc subsystem decision process as a step response. In practice, hormonal shifts between leptin and ghrelin are coupled with other signals like gastric distension to inform this decision process on food intake [Ferrario *et al.* 2016; Farr *et al.* 2016]. Future work should experiment with varying functions of heightened food reward (i.e., used as a proxy of OFC-Amygdala-NAcc subsystem evaluation) and compare its results to this work.

Another avenue of future work could reimagine the incorporation of interoceptive information. Beukman *et al.* [2023] challenge the classical integration of contextual information in a CMDP suggesting its conflation of context and state information to an RL agent. Future work could utilise their proposed Decision Adapter architecture to incorporate information related to ghrelin levels. It would be interesting to compare if additional factors like heightened food reward, food location and hunger thresholds would demonstrate the same inter-relations and influence on the RL agent's behaviour or if these parameters would be necessary at all.

Lastly, an RL agent incorporating interoception provides a simple platform to test neuroscientific theories related to food intake and the dysregulation of its management associated with medical conditions [Murray *et al.* 2021; Timper and Brüning 2017; Stover *et al.* 2023]. Focusing on a specific ailment like depressive disorders and manipulating characteristic of the MiniGrid environment, food location and food reward could help better understand the effect of these conditions on an organism's behaviour.

## 5.3 Concluding remarks

The results in Chapter 4 show that interoception can be incorporated into an RL pipeline by considering interoceptive information as contextual information (i.e., assuming a Contextual Markov Decision Process) and integrating it through methods used by Li [2017] and Ball *et al.* [2021]. The results obtained demonstrate that equipping an RL agent with interoception influences the agent's decision making and behaviour as observed in living organisms [Tsakiris and Critchley 2016] by utilising a baseline and heightened food reward function to recreate the holistic valuation process done by the OFC-Amygdala-NAcc subsystem in the brain detailed in Chapter 2. This is demonstrated in the influence of interoception on an agent's learn policy (i.e., always intersecting the food location) and unconventional behaviour to respond to interoceptive needs (i.e., leaving a goal state to "eat" at a food location). With varying key parameters of hunger threshold, food location and heightened food reward, the RL model which incorporates interoception is able to recreate natural survival-based behaviours observed in a real-life mouse. In alternative settings, such as an RL agent possessing internal ghrelin control actions or a proportional food reward function, food location and hunger threshold do not affect the agent's learned policy. When these parameters are set to extremes (i.e., extremely low heightened food reward, low hunger threshold or distant food location), the RL model which incorporates interoception exhibits dysregulated feeding behaviours associated with medical conditions like depressive anhedonia and obesity [Zigman *et al.* 2016; Coccurello 2019]. Such models can be used to investigate theories and behaviours related to motivated behaviours and such clinical conditions.

# Appendix A

# Appendix 1: Hyperparameters

| **Hyperparameter** | **Setting** |
|---|---|
| Number of episodes | 15,000 |
| Replay memory size | 100,000 |
| Batch size | 128 |
| Epsilon start | 1 |
| Epsilon end | 0.05 |
| Epsilon decay step | $1 \times 10^{-6}$ |
| Learning rate | $1 \times 10^{-3}$ |
| Tau | 0.01 |
| Gamma | 0.99 |
| Beta | 1 |

Table A.1: Hyperparameter settings

# Appendix B

# Appendix 2: Complete circuits by Averbeck and Murray [2020]

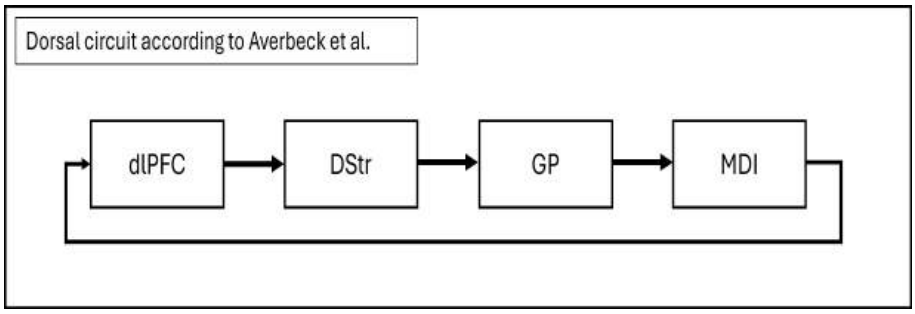



Figure B.1: Complete dorsal circuit described by Averbeck and Murray [2020]

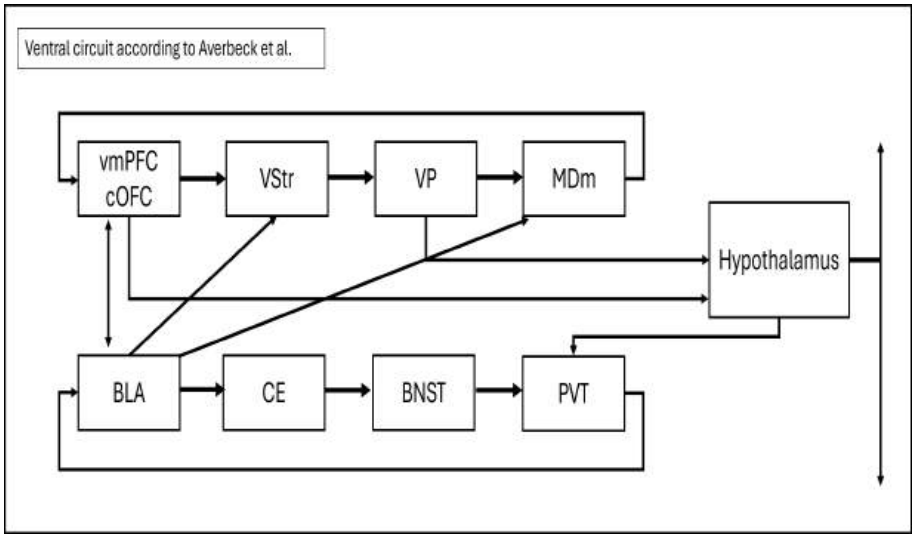



Figure B.2: Complete ventral circuit described by Averbeck and Murray [2020]

# Appendix C

# Appendix 3: List of Acronyms

- LH – Lateral hypothalamus
- ARC – Arcuate nucleus
- VMN – Ventromedial nuclei
- NAcc – Nucleus Accumbens
- VTA – Ventral Tegmental Area
- PFC – Prefrontal cortex
- mPFC – Medial Prefrontal cortex
- OFC - Orbitofrontal cortex
- dlPFC – Dorsolateral prefrontal cortex
- vlPFC – ventrolateral PFC
- vmPFC – ventromedial PFC
- ACC – Anterior Cingulate cortex
- dACC - Dorsal Anterior cingulate cortex
- pre-SMA – Presupplementary Motor Area

# References


[Abdalla 2015] Mona Mohamed Ibrahim Abdalla. Ghrelin–physiological functions and regulation. *European endocrinology*, 11(2):90, 2015.

[Aubret *et al.* 2019] Arthur Aubret, Laetitia Matignon, and Salima Hassas. A survey on intrinsic motivation in reinforcement learning. *arXiv preprint arXiv:1908.06976*, 2019.

[Averbeck and Murray 2020] Bruno B Averbeck and Elisabeth A Murray. Hypothalamic interactions with large-scale neural circuits underlying reinforcement learning and motivated behavior. *Trends in neurosciences*, 43(9):681–694, 2020.

[Baik 2021] Ja-Hyun Baik. Dopaminergic control of the feeding circuit. *Endocrinology and Metabolism*, 36(2):229–239, 2021.

[Ball *et al.* 2021] Philip J Ball, Cong Lu, Jack Parker-Holder, and Stephen Roberts. Augmented world models facilitate zero-shot dynamics generalization from a single offline environment. In *International Conference on Machine Learning*, pages 619–629. PMLR, 2021.

[Barto *et al.* 2004] Andrew G Barto, Satinder Singh, Nuttapong Chentanez, et al. Intrinsically motivated learning of hierarchical collections of skills. In *Proceedings of the 3rd International Conference on Development and Learning*, volume 112, page 19. Citeseer, 2004.

[Barto 2012] Andrew G Barto. Intrinsic motivation and reinforcement learning. In *Intrinsically motivated learning in natural and artificial systems*, pages 17–47. Springer, 2012.

[Beukman *et al.* 2023] Michael Beukman, Devon Jarvis, Richard Klein, Steven James, and Benjamin Rosman. Dynamics generalisation in reinforcement learning via adaptive context-aware policies. *Advances in Neural Information Processing Systems*, 36:40167–40203, 2023.

[Chau *et al.* 2018] Bolton KH Chau, Huw Jarvis, Chun-Kit Law, and Trevor T-J Chong. Dopamine and reward: a view from the prefrontal cortex. *Behavioural pharmacology*, 29(7):569–583, 2018.

[Chentanez *et al.* 2004] Nuttapong Chentanez, Andrew Barto, and Satinder Singh. Intrinsically motivated reinforcement learning. *Advances in neural information processing systems*, 17, 2004.

[Chevalier-Boisvert *et al.* 2018] Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minimalistic gridworld environment for openai gym (2018). *URL https://github. com/maximecb/gym-minigrid*, 6, 2018.

[Coccurello 2019] Roberto Coccurello. Anhedonia in depression symptomatology: Appetite dysregulation and defective brain reward processing. *Behavioural Brain Research*, 372:112041, 2019.

[Date *et al.* 2002] Yukari Date, Noboru Murakami, Koji Toshinai, Shigeru Matsukura, Akira Niijima, Hisayuki Matsuo, Kenji Kangawa, and Masamitsu Nakazato. The role of the gastric afferent vagal nerve in ghrelin-induced feeding and growth hormone secretion in rats. *Gastroenterology*, 123(4):1120–1128, 2002.

[Erdmann *et al.* 2003] Johannes Erdmann, Florian Lippl, and Volker Schusdziarra. Differential effect of protein and fat on plasma ghrelin levels in man. *Regulatory peptides*, 116(1-3):101–107, 2003.

[Farr *et al.* 2016] Olivia M Farr, R Li Chiang-shan, and Christos S Mantzoros. Central nervous system regulation of eating: Insights from human brain imaging. *Metabolism*, 65(5):699–713, 2016.

[Ferrario *et al.* 2016] Carrie R Ferrario, Gwenaël Labouèbe, Shuai Liu, Edward H Nieh, Vanessa H Routh, Shengjin Xu, and Eoin C O'Connor. Homeostasis meets motivation in the battle to control food intake. *Journal of Neuroscience*, 36(45):11469–11481, 2016.

[Floresco 2015] Stan B Floresco. The nucleus accumbens: an interface between cognition, emotion, and action. *Annual review of psychology*, 66(1):25–52, 2015.

[Gordan and Krishanan 2014] Marzieh Gordan and Isai Amutan Krishanan. A review of bf skinner's 'reinforcement theory of motivation'. *International journal of research in education methodology*, 5(3):680–688, 2014.

[Greenwood *et al.* 2011] Benjamin N Greenwood, Teresa E Foley, Tony V Le, Paul V Strong, Alice B Loughridge, Heidi EW Day, and Monika Fleshner. Long-term voluntary wheel running is rewarding and produces plasticity in the mesolimbic reward pathway. *Behavioural brain research*, 217(2):354–362, 2011.

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

[Heinrichs 2001] Stephen C Heinrichs. Mouse feeding behavior: ethology, regulatory mechanisms and utility for mutant phenotyping. *Behavioural brain research*, 125(1-2):81–88, 2001.

[Jensen *et al.* 2013] Thais L Jensen, MK Kiersgaard, Dorte B Sørensen, and LF Mikkelsen. Fasting of mice: a review. *Laboratory animals*, 47(4):225–240, 2013.

[Kakade and Dayan 2002] Sham Kakade and Peter Dayan. Dopamine: generalization and bonuses. *Neural Networks*, 15(4):549–559, 2002.

[Kayal *et al.* 2025] Aya Kayal, Eduardo Pignatelli, and Laura Toni. The impact of intrinsic rewards on exploration in reinforcement learning. *Neural Computing and Applications*, pages 1–35, 2025.

[Kelley *et al.* 2005] Ann E Kelley, Brian A Baldo, Wayne E Pratt, and Matthew J Will. Corticostriatal-hypothalamic circuitry and food motivation: integration of energy, action and reward. *Physiology & behavior*, 86(5):773–795, 2005.

[Kim 2013] Sung-il Kim. Neuroscientific model of motivational process. *Frontiers in psychology*, 4:98, 2013.

[Latyshev and Panov 2024] AK Latyshev and Aleksandr I Panov. Intrinsic motivation in model-based reinforcement learning: A brief review. *Scientific and Technical Information Processing*, 51(5):460–470, 2024.

[Lee *et al.* 2020] Kimin Lee, Younggyo Seo, Seunghyun Lee, Honglak Lee, and Jinwoo Shin. Context-aware dynamics model for generalization in model-based reinforcement learning. In *International Conference on Machine Learning*, pages 5757–5766. PMLR, 2020.

[Levine and Billington 1997] Allen S Levine and Charles J Billington. Why do we eat? a neural systems approach. *Annual review of nutrition*, 17(1):597–619, 1997.

[Lewis *et al.* 2021] Robert G Lewis, Ermanno Florio, Daniela Punzo, and Emiliana Borrelli. The brain's reward system in health and disease. In *Circadian Clock in Brain Health and Disease*, pages 57–69. Springer, 2021.

[Li 2017] Yuxi Li. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*, 2017.

[Liu and Kanoski 2018] Clarissa M Liu and Scott E Kanoski. Homeostatic and non-homeostatic controls of feeding behavior: Distinct vs. common neural systems. *Physiology & behavior*, 193:223–231, 2018.

[Merrick 2012] Kathryn E Merrick. Intrinsic motivation and introspection in reinforcement learning. *IEEE Transactions on Autonomous Mental Development*, 4(4):315–329, 2012.

[Mnih *et al.* 2013] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[Mohapatra and Wagner 2023] Alok Nath Mohapatra and Shlomo Wagner. The role of the prefrontal cortex in social interactions of animal models and the implications for autism spectrum disorder. *Frontiers in Psychiatry*, 14:1205199, 2023.

[Murray *et al.* 2014] Susan Murray, Alastair Tulloch, Mark S Gold, and Nicole M Avena. Hormonal and neural mechanisms of food reward, eating behaviour and obesity. *Nature Reviews Endocrinology*, 10(9):540–552, 2014.

[Murray *et al.* 2021] Susan M Murray, Carina S Brown, Walter H Kaye, and Christina E Wierenga. Anhedonia in eating disorders. In *Anhedonia: Preclinical, Translational, and Clinical Integration*, pages 219–236. Springer, 2021.

[Narkhede *et al.* 2022] Meenal V Narkhede, Prashant P Bartakke, and Mukul S Sutaone. A review on weight initialization strategies for neural networks. *Artificial intelligence review*, 55(1):291–322, 2022.

[Nian *et al.* 2020] Rui Nian, Jinfeng Liu, and Biao Huang. A review on reinforcement learning: Introduction and applications in industrial process control. *Computers & Chemical Engineering*, 139:106886, 2020.

[Nieh *et al.* 2016] Edward H Nieh, Caitlin M Vander Weele, Gillian A Matthews, Kara N Presbrey, Romy Wichmann, Christopher A Leppla, Ehsan M Izadmehr, and Kay M Tye. Inhibitory input from the lateral hypothalamus to the ventral tegmental area disinhibits dopamine neurons and promotes behavioral activation. *Neuron*, 90(6):1286–1298, 2016.

[Perlman 2016] Robert L Perlman. Mouse models of human disease: an evolutionary perspective. *Evolution, medicine, and public health*, 2016(1):170–176, 2016.

[Rathod and Di Fulvio 2021] Yakshkumar Dilipbhai Rathod and Mauricio Di Fulvio. The feeding microstructure of male and female mice. *PloS one*, 16(2):e0246569, 2021.

[Rolls 2023] ET Rolls. The orbitofrontal cortex, amygdala, reward value, emotion, and decision-making. *Brain computations and connectivity. Oxford University Press. https://doi. org/10.1093/oso/9780198887911.003*, 11, 2023.

[Ruder 2016] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[Saper and Lowell 2014] Clifford B Saper and Bradford B Lowell. The hypothalamus. *Current Biology*, 24(23):R1111–R1116, 2014.

[Simpson and Balsam 2015] Eleanor H Simpson and Peter D Balsam. The behavioral neuroscience of motivation: an overview of concepts, measures, and translational applications. *Behavioral neuroscience of motivation*, pages 1–12, 2015.

[Skibicka *et al.* 2011] Karolina P Skibicka, Caroline Hansson, Mayte Alvarez-Crespo, P Anders Friberg, and Suzanne L Dickson. Ghrelin directly targets the ventral tegmental area to increase food motivation. *Neuroscience*, 180:129–137, 2011.

[Stover *et al.* 2023] Patrick J Stover, Martha S Field, Mark L Andermann, Regan L Bailey, Rachel L Batterham, Elizabeth Cauffman, Gema Frühbeck, Per O Iversen, Pamela Starke-Reed, Scott M Sternson, et al. Neurobiology of eating behavior, nutrition, and health. *Journal of internal medicine*, 294(5):582–604, 2023.

[Sutton *et al.* 1998] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[Timper and Brüning 2017] Katharina Timper and Jens C Brüning. Hypothalamic circuits regulating appetite and energy homeostasis: pathways to obesity. *Disease models & mechanisms*, 10(6):679–689, 2017.

[Tsakiris and Critchley 2016] Manos Tsakiris and Hugo Critchley. Interoception beyond homeostasis: affect, cognition and mental health. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1708):20160002, 2016.

[Zigman *et al.* 2016] Jeffrey M Zigman, Sebastien G Bouret, and Zane B Andrews. Obesity impairs the action of the neuroendocrine ghrelin system. *Trends in Endocrinology & Metabolism*, 27(1):54–63, 2016.