Obesity and environmental uncertainty

Luis Nicolás Luarte Rodríguez

Introduction

[—]

Food-seeking behavior and uncertainty

Uncertainty implies situations where there is partial or incomplete information. When considering agency in decision-making tasks, uncertainty, refers to the incomplete information about the outcome that a given decision will generate, and also incomplete information about the probability distribution governing such outcome, whereas 'risk' implies knowledge about such probability distribution (De Groot & Thurik, 2018). A particular feature of risk is the inverted 'U' shape, where at outcome probability 0 or 1 is at minima, and at maxima when probability is 0.5, this derives from risk being measured as outcome variance (Preuschoff et al., 2006). If an agent is situated in a natural environment, it is unlikely to have complete information about the decision-outcome pairing, and thus is forced to generates estimates or more complex models about environment statistical properties.

When considering the case of an agent in a natural environment, one of the most relevant cases is that of foraging. Foraging implies an agent searching for resources in a partially known environment, with depleting resources, and the supposition that the agent seeks to maximize its resources in the most efficient possible way, while accounting for unknown resource distribution (Charnov, 1976). Multiple theories on how an agent must decide to optimally allocate time to each resource patch (Wajnberg et al., 2000), optimal path (Hills et al., 2013; Humphries & Sims, 2014) and evolutionary roots (Wosniack et al., 2017). However, how environmental variables, such as resources uncertainty, modulate agent decision making have been less reviewed. Intuitively, if an agent does not have perfect information about the environment to be able to predict if there is going to be food in the future, it should consider a proxy of this, such as current food availability, which integrated over a time span, indicates the level of uncertainty of food availability.

To consider the modulation of uncertainty in decision making, one needs to take into account that uncertainty in food availability is a direct product of food scarcity, as the probability of food encounters is reduced proportional to scarcity. Thus, is to be expected

that uncertainty effects are in line with food scarcity signaling, that is, a expected reduction in energy expenditure in order to preserve the energetic balance. However empirical evidence is not so clear in this regard (Polo, 2002), as both increases and reductions in body mass given increased levels of uncertainty in food availability (Fokidis et al., 2012). When controlling for total food intake in condition of fixed or variable food availability, studies with birds show a decrease in body mass, which is explained, in part, by increased locomotor activity (Fokidis et al., 2012), this results are discussed in terms of stress (augmented due to food variability). However they can also be interpreted of expected foraging behavior.

When an environment presents higher uncertainty on food availability, it could be expected that food-seeking bouts are to be increased, in order to compensate for reduced food encounters or, complementary to that, an increased hoarding-type behavior. In the face of uncertainty, specifically, regarding food resources, a typical behavior is to increase food-seeking bouts, and resulting hoarding-type behavior, allowing the consumption of extra calories. This has been considered as a mechanism to prevent possible starvation because uncertainty is used as a proxy of future food scarcity, so eating in excess could prevent starvation (Anselme & Güntürkün, 2019). However, this hoarding-behavior can also be explained by directly estimating food availability in the environment. Food scarcity can generate multiple change in number of foraging spots visited, diet diversity, and others (Harris et al., 2010), which are similar to the ones already discussed. Nevertheless, scarcity does not imply uncertainty, because the caloric density of an environment can decrease, but food availability, while less, may remain constant. When previous regularities regarding feeding routine, such as feeders positions are constantly changed, increased intake is observed relative to the unchanging environment (Forkman, 1993). This shows that, while food scarcity could be a cause of hoarding behavior, environment properties are enough to trigger such behaviors. In the previous case, environmental condition and food reserves were altered. However, its levels were more than sufficient to satisfy energetic demands. Nevertheless, the uncertainty effect on food intake holds even when food levels are equated (through equating total time available to get food) across predictable and unpredictable settings (Cuthill, 2000). In addition to increased intake, under uncertain conditions, mathematical modelling of foraging behavior, shows that an optimal strategy is to change foraging bout to the start of the day, in order to account for missed foraging bouts (Bednekoff & Houston, 1994).

Up to this point, there are at least 3 points to consider regarding food uncertainty, (1) under food availability uncertainty, food-seeking behavior is increased (Fokidis et al., 2012; Polo, 2002; Robinson et al., 2014). (2) when uncertainty is increased strategies to maintain energetic balance, such as hoarding (Anselme & Güntürkün, 2019) or increasing body mass (Cuthill, 2000; Moiron et al., 2018) emerge. (3) such strategies imply a trade-off between preventing starvation and increasing risk of predation (due to reduced mobility),

and as such are suggestive of a dynamic balance (Macleod et al., 2005).

To address food-seeking behavior under uncertainty, one needs to consider which behavior is reflective of a food-seeking action. In typical experimental settings, such action is reflected by interaction with cues or apparatuses that are related to reward delivery, when the action is interacting with a conditioned stimuli is called sign-tracking, whereas if interaction is with food dispenser is called goal-tracking (Silva et al., 1992). More specifically, sign-tracking, refers to an approaching behavior towards previously conditioned stimuli and rewards. So, it implies a previous conditional-stimulus and unconditional stimulus pairing, and, afterwards, tracking of the signal that was previously associated with the reward (Flagel, 2014).

When uncertainty is introduced at the stage of conditional and unconditional stimulus pairing, as the probability of reward delivery upon lever pressing, sign-tracking increases as the probabilities of reward delivery approaches 50%, and the amount of reward is more varied (Anselme et al., 2013). In this case, as the delivery of a given reward gives no information about following one (delivery is determined by a probability function, independent of animal action), it can be assumed that, under Shannon entropy formulation, entropy (which can be understood as a measure of uncertainty) (Namdari & Li, 2019) reaches the peak at 50% probability, and, furthermore, it predicts that uniform distributions, with more outcomes, increase uncertainty. Both were the case in the previously presented experiment (assuming uncertainty drove signal-tracking) as 50% probability of delivering 2 or 0 pellets had lower signal-tracking than 50% probability of delivering 0 or 1, 2, or 3 pellets with equal probability (16.7% for 1, 2 or 3 pellets). This, again, points out that increased food-seeking related behavior increases upon increased uncertainty even when food availability is controlled. This effect has been replicated in studies with amphetamine sensitization, where uncertainty (on conditioned stimulus and unconditioned stimulus) and sensitization, independently, augmented sign-tracking behavior, however, the effect of both uncertainty and sensitization was not additive suggesting a ceiling effect (Robinson et al., 2015).

The increased, food-seeking related behavior magnification by uncertainty, has been found with partial reinforcement procedures (Collins et al., 1983), with manipulation of food placement variability (Forkman, 1993), variability on reward quality and delivery delay (Craft, 2016), and in sequential probability tasks (Stagner & Zentall, 2010). Implying a robust effect across multiple food-related uncertainty scenarios.

Assessing and dealing with uncertainty

From the perspective of a foraging animal, food sources are distributed in a partially known space, where effort must be made to obtain such sources. Uncertainty, reveals the consistency of food sources in a given space, where more uncertainty determines more

difficulty in obtaining food. However, the consistency of food sources must be sensed through a mechanism that updates its estimate in a trial by trial basis, because is safe to assume that an agent interested in sensing environment uncertainty does not possess complete information. A plausible mechanism is to sense uncertainty, indirectly, via the reward prediction error. The reward prediction error is simply

$$actual\,reward - expected\,reward$$

As the reward prediction error is thought to operate in environments where a particular action lead to a probable reward, this error is used to update the value of any given action, then, the value of such time step (which can be associated with a given action) is given by the discounted rewards from that point onwards up to the termination of the trial series (Sutton & Barto, 2018)

$$expected\,reward = reward_{t+1} + \gamma reward_{t+2} + \gamma^2 reward_{t+3} + \ldots + \gamma^k reward_T$$

Here the trial series is composed of T time steps with a discount factor $\gamma, 0 \leq \gamma \leq 1$. The discount factor is there to signal the typical preference for obtaining rewards now rather than latter, on and how big it is will depend on properties of both agent and environment (Glimcher, 2011).

The formulation presented above is just a mathematical representation of several assumptions of how an agent can learn expected reward values in a finite, trial based, experiment and then calculate the prediction error at each time step, how this reward prediction error is used to update values will be presented latter on. However, the main idea is that over trials, as the expected value approximates the real one, the reward prediction error goes down, nevertheless, if rewards value change the error goes up reflecting this change (see figure 1).

The main neural circuitry supporting the computation of the reward prediction error is thought to be supported, mainly, by the dopamine system (Schultz, 2016). As the reward prediction error was first derived from behavioral data, to assess the biological feasibility three components must exist (1) expectation encoding units; (2) reward encoding units and (3) a subtraction unit (Watabe-Uchida et al., 2017)

Reward prediction errors models predict both three cases (1) where the expected reward and current reward are equal (no prediction error); (2) expected reward is less than the current reward (negative error) or (3) expected reward is greater than current reward (positive error). Midbrain dopamine neurons have been found to encode positive error but not negative under reinforcement learning models (Bayer & Glimcher, 2005). Around this point two main hypothesis have been formulated, the first, proposes than negative error are encoded via lowering the fire-rate compared to the baseline (Schultz et al., 1997),

whereas the second, proposes an opponency system between dopamine and serotonin systems (Daw et al., 2002). Dopamine neurons in the ventral tegmental area have been found to encode the future discounted rewards (Enomoto et al., 2011). This two lines of evidence points that dopamine is capable of encoding expectation, reward value and doing subtraction (perhaps including the serotonin system), showing a significant complexity of this system, which might exceed value-related computations (Takahashi et al., 2017).

Above the function of dopamine neurons in reward prediction error has been stated, more specifically, this function seems to be related to the phasic activations, whereas, more sustained activation is related to reward uncertainty (measured as reward variance, thus reaching its peak at a probability of 0.5) (Fiorillo, 2003). Such uncertainty-related signal has also been found in the orbito frontal cortex, amygdala (Schultz et al., 2008) and medial frontal lobe (Huettel, 2005). A plausible hypothesis to link the reward prediction error and uncertainty encoding, can state that, over time, reward prediction error signals are integrated into an uncertainty signal, as, over time, more error is to be expected under higher reward variability (see figure 2). However, evidence points towards independent signals of reward prediction error and uncertainty in the orbito frontal cortex (Rushworth & Behrens, 2008). Nevertheless, at least at a computational level, the reward prediction error can be used to estimate the reward-related uncertainty (Soltani & Izquierdo, 2019).

As seen previously the reward prediction error is linked with uncertainty, as the error is a function over uncertainty levels. The way in that these two are connected can be seen in a more contextualized fashion. The main purpose of reward prediction error is to allow an agent to assess the value of actions in a certain environment, while measuring how stable is such environment. Thus, reward prediction error assists in learning, and in such context it serves the purpose of regulating the learning rate. Learning rate is parameter that defines how, a reward prediction error (or any measure of error), should affect posterior decisions, intuitively, when starting in a given task, higher learning rates are to be expected so learning happens at faster rate, however, as the task is properly learned learning rates should go down, so to not be influenced by random fluctuations (Even-Dar & Mansour, 2001). Mathematical models, based on previously presented dopamine research, have proposed that learning rates are to be updated via the covariance between predictions (expected rewards) and prediction errors (Preuschoff & Bossaerts, 2007), in the same vein, empirical experiments have shown that humans behave according to a, reward standard deviation-dependent scaling of reward prediction error (Diederen & Schultz, 2015), so error should be less impactful when standard deviation is high. This proposed learning rate modifications are in line with the original model by Pearce & Hall (1980), which proposed that 'surprise' affected the learning rate, or viewed from the other side, as the pairing between unconditioned stimulus and conditioned stimulus became more predictable, the 'associability' decreased. Is important to note that other systems, aside from dopamine, are able to track environment uncertainty level, such as the endocrine system through the

stress response, measured as subjetive stress, pupil diameter and skin conductance (de Berker et al., 2016).

Changing the learning rate based on the reward prediction error, reflects a constant tension an agented, faced with an uncertain environment, must face. How new acquired information must be considered?, a notion to answer this question is that of expected and unexpected uncertainty (Yu & Dayan, 2005). Expected uncertainty is the variability attributable to the stochastic nature of the reward, whereas unexpected uncertainty assumes that the agent is creating belief about action-rewards associations, and incoming information breaks such beliefs (Payzan-LeNestour et al., 2013). Unexpected uncertainty, thus, has been proposed as top-down process which might be tracked by the Locus Coeruleus norepinephrine activity (Filipowicz et al., 2020; Payzan-LeNestour et al., 2013). Moreover, such activity, measured as pupil diameter, has been found to track the learning rates (Nassar et al., 2012), which represents the end product of assessing environment in terms of expected and unexpected uncertainty. The intuition here is that when the statistical properties of an environment are changed, this should generate a signal of unexpected uncertainty similar to surprise as proposed by Pearce & Hall (1980), which in turn represent that the current model of the environment must change in order to accurately depict it. Then, the learning rate must increase, so to give more weight to more recent information and correctly update the environment model (Faraji et al., 2017).

Up until know the discussion presented has focused on the calculation of reward value, typically, given a certain action. However, the computation of action-reward value is not directly linked to action choice. Picture a situation where the environment present high levels of uncertainty, and one action, until the present time step, have been associated with high rewards. If we were to choose just based on the maximum reward in such environment, we could miss potential better options, which true value, cannot be appropriately calculated because of variability. Such situation is more specifically defined by the exploration/exploitation dilemma, which posits that an agent, in order to obtain rewards, must 'exploit' current knowledge. However, it also must 'explore' to determine the best option in the future (Sutton & Barto, 2018). One of the findings is that the manipulation of dopamine levels modulates striatal representation of reward prediction errors, and subjects treated with L-dopa (a metabolic precursor of dopamine) chose, with more frequency, the option with greater reward compared to the placebo group and haloperidol (dopamine receptor antagonist) group (Pessiglione et al., 2006). Although the authors did not report the temperature parameter (the one that determines the balance between exploration and exploitation) of the model, given that in all three group the optimal option was learned, it can be interpreted that the increase in dopamine levels effectively induced a bias toward exploitation. Direct evidence on the effect of L-dopa in exploration/explotation parameters, effectively shows that exploration is reduced, and this is associated with modulation of uncertainty signals in the insula and anterior cingulate

cortex (Chakroun et al., 2020).

Taking action in uncertain environments

Previously, the notion that there is something linking the estimated values and the actions taken was presented in terms of exploration/exploitation. How an agent decides, based on its estimated, to behave at any given time is called a 'policy', and as such, it constitutes a mapping from estimates and actions (Sutton & Barto, 2018). As presented previously, the reward prediction error representation is able to guide the chosen policy of the agent (Pessiglione et al., 2006). Heuristics, which are strategies that rely heavily on exploiting environment statistical properties, have been proposed to be guiding decision-making in uncertain environments (Hafenbrädl et al., 2016). Some heuristics are thought to be an evolutionary derivative from uncertain environments (van den Berg & Wenseleers, 2018). However, what aspect of uncertainty is the one used to selected the optimal policy is not clear (Gershman, 2018). Gershman (2018) explored the fit of two models to human behavioral data in two-armed bandit tasks, and found signatures of both models in behavioral data, while a mixture of both models more salient signatures represented the best fit. The first model fit corresponded to Upper-Confidence-Bound (Auer et al., 2002), the intuition is that an agent should choose based on the times a certain action has been taken, and the potential value of each action on the environment. The action selection is formally assigned as:

$$A_t := \underset{a}{\operatorname{argmax}} \left[Q_t(a) + c\sqrt{\frac{ln_t}{N_t(a)}} \right]$$

Here $Q_t(a)$ represent the expected value of taking action a, in $c\sqrt{\frac{ln_t}{N_t(a)}}$ the denominator represents how many times action a has been chosen up to a certain point t, as time ln_t appears in the numerator, when action a is not chosen its upper-bound will increase, but decrease if such is continously chosen. Note that Gershman (2018) used a modified version of this algorithm to reflect human decision stochasticity, nevertheless, this provides enough insight into how it considers uncertainty. Author found indirect support for this policy, by considering that reaction times are faster when estimated rewards are more different (Tajima et al., 2016), and that reaction time decreased in proportion to increasing relative uncertainty, thus acting according to an uncertainty bonus as posed from the Upper-Confidence-Bound. The second model examined corresponded to Thompson Sampling, which builds reward priors on each option, this priors are beta-distributed with parameters α and β according with the formula:

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},\ \theta \in [0,1]$$

Where θ is the model expected reward, and Γ represents the Gamma function.To illustrate how the priors are update a win/loss reward environment can be considered. First, the agent will sample from each of the distribution, and will choose the action associated with the distribution that gave the largest sample. Then, if the reward is a win (or '1') the α is updated as $\alpha = \alpha + 1$ and $\beta = \beta$, when the observed reward is a loss (or '0') $\beta = \beta + 1$ and $\alpha = \alpha$. Findings by Gershman (2018) noted that signatures corresponding to this model were that choice stochasticity (exploration action) were proportional with the level of uncertainty in the model distributions.

Evidence for directed exploration, based on uncertainty levels, has been found in humans (Blanco & Sloutsky, 2019; Wilson et al., 2014). However, while directed exploration seems to be a robust strategy in humans, certain aspects emerge and vary throughout the life span (Somerville et al., 2017), pointing towards a complex and dynamic system. The main idea behind the previously presented models are twofold (1) uncertainty, in someway, guides the balance between exploration and exploitation and (2) simple computations can sufficiently describe exploratory behavior under varying levels of uncertainty. Integrating over evidence of foraging under uncertainty and computational models presented, food-seeking behavior can be stated as a series of actions, occurring in an uncertain environment, where each action (feeding bout) is evaluated in terms of the reward prediction error. Reward prediction error, however, not only informs about the value of the expected and current reward, but also, by considering the history of such errors, calculates environmental uncertainty (Mikhael & Bogacz, 2016) (see figure **??**) and represents it at a neural level (Fiorillo, 2003). Finally, action policies are intimately related to uncertainty, thus establishing a clear link between feeding-related behavior and environment uncertainty.

Uncertainty representations at the neural level

If uncertainty can modulate food-seeking behavior in order to increase intake and better sustain energetic reserves, it is expected to have at least to functional instances (1) an uncertainty sensing unit and (2) a reward processing unit, which can relay information to homeostatic-related and decision-making loci, to integrate such information and determine the next action to take. To determine the neural substrates of such instances, environment-agent dynamics can be represented through Markov decision tasks. Such tasks consider a set of states with possible transitions between each one, and two functions: (1) the one in charge of determining the state transition given the agent action and (2) an action-state-reward function, which maps a reward to a given action-state tuple. In such tasks, uncertainty is derived from probability matrices assigned to either of the two functions. When state transition functions are manipulated, two scenarios can be created: (1) a regular one, where action-state transitions are deterministic, and (2) a random one, where action-state can not be predicted.

A plausible neuronal substrate underlying such functions is the striatum, as it has been found to be involved in decision-making actions such as action selection (Balleine et al., 2007), action value representation (Samejima, 2005), and reward representation (Wake & Izuma, 2017). Markov decision tasks can be used to test behavior under certain or uncertain environments. Such task comprises of a set of states, which are accessible by performing certain actions, however, the function that defines the mapping between actions and states can be random or deterministic. Finally, entering a given state provides a reward. When comparing learning of certain versus uncertain Markov decision tasks, the dorsal striatum seems to be more associated with the uncertain condition, whereas the dorso lateral prefrontal cortex showed greater activation under the certain condition (Tanaka et al., 2006). This can be explained in terms of immediate and long-term reward prediction, as state transitions are more uncertain, subjects can only reliably predict the following rewards, in turn, if state-transition dynamics are deterministic, the reward over a long series of actions can be predicted, thus, making useful to consider rewards in the long term. A similar relationship between uncertainty and immediate/long-term rewards has been noted in human consumers when exposed to features typically associated with environmental uncertainty, such as, economic crisis, unemployment, among others (van den Berg & Wenseleers, 2018).

If an environment is stable, then state-action-reward mappings can be optimized to reduce reward-prediction error. In this way, when the mapping is optimized, reward-related circuitry should reduce its activity (Friston, 2009). However, this mapping is always modulated by environment dynamics regarding uncertainty. An optimal mapping in a given environment state can increase the reward prediction error in the same environment if this is non-stationary. The anterior cingulate cortex (ACC) has been shown to increase its activation levels when predictability in the environment drops (Davis et al., 2010), effectively signaling environment dynamics.

As previously stated, environment dynamics need to be taken into account in order to appropriately interpret obtained rewards. If I visit a restaurant and the food served is delicious, my rating of the restaurant should not be too hasty as this could be just good luck. However, if this has always been the case, giving a high rating would be the correct choice. In term of rewards, uncertainty is high when a given rewards give no information about the ones to come, conversely, certainty is achieved when a given reward gives all information about the following one. Direct tracking of environment volatility has been found to be well represented in the ACC (Behrens et al., 2007), presumably by encoding some sort of learning rate that bias valuation of rewards more to the short-term if volatility is high, and to the long-term is volatility is lower. The competing hypothesis of ACC describes its function to a decision-difficulty sensing unit, or demand of control when overriding default action is more optimal (Shenhav et al., 2016). However, it should be noted that Behrens et al. (2007) results were circumscribed to the time point where the

outcome is observed, which corresponds to the proper timing to assign obtained reward influence to the following behavior.

When representing the uncertainty of a given environment, an agent must pair the value obtained with the action performed. For each action possible, the agent updates the value of the action-reward tuple based on the reward prediction error.

Temporal dynamics of action-reward pairing and reward prediction error are such that the former occurs first relative to the later. Such temporal difference is reasonable because the pairing should be represented when taken action, and the prediction error requires feedback in order to compare obtained versus expected rewards. Considering this, the action-reward pairing has been found to be correlated to activity at the putamen, whereas rewards-prediction error, to be represented in the caudate nucleus (Haruno & Kawato, 2006). However, as the authors point, both structures are likely to be involved in a larger loop containing the ACC, which would make sense to integrate reward evaluation over states, actions, and environmental uncertainty, and optimally influence following behavior.

It can be inferred from the way action-reward pairing is stated that it corresponds to action selection based on a history of rewards, which are mediated by the reward prediction error. Inhibition of putamen activity has effectively shown a reduction in performance when the task requires the consideration of reward history to select correct actions (Muranishi et al., 2011). Signal encoding, however, seems to be more complex, as basal ganglia direct pathway encode rewards outcomes, and the indirect pathway represents the next-action selection (Nonomura et al., 2018). Together, this points to a multi-structure network that represents expected and obtained rewards as an error, which allows easing computational requirements as the current state needs only to be compared with the expectation, that encompasses all previous history of rewards. Moreover, this signal updates rewards given actions, while considering environment volatility and the proper weighting of immediate versus long-term rewards. Thus, allowing to optimize behavior even when environments are non-stationary and rapidly changing.

Orexin/Hypocretin system and stress in food-seeking behavior (To Do)

Up to this point, the way reward-related systems interact with environmental uncertainty has been discussed. Several structures seem to be involved in integrating reward value in the face of environment volatility. Moreover, empirical findings of food-seeking behavior in predictable/unpredictable environments were pointed out. However, the direct mechanism that guides food-seeking behavior is lacking. One such system is the Orexin/Hypocretin (HO), which is part of the energetic homeostasis and feeding pathways (Toshinai et al., 2003), playing a large role in increasing food intake (Wolf, 2009). However, a more broad and complex opioid system is thought to control food intake, which in turn is modulated

by food preference, and has proven to be selective to certain macro-nutrients, such as fat (Taha, 2010). More recent evidence has linked the activation of the hypothalamic HO system to an increase in short-term spatial memory, which is a function that supports exploratory foraging behavior (Aitta-aho et al., 2016).

Moreover, orexin promotion of such foraging-related behavior has been postulated as one of its main functions (Barson, 2020). Such function is relevant because foraging behavior evolved in a specific type of environment, where resources are sparse, clustered, and is a potential risk of predation, and developed relatively stable strategies to deal with such conditions (Wosniack et al., 2017). Thus, foraging behavior seeks to generate a strategy to maximize energetic intake in a partially known environment. However, if environment resources are non-depleting, it can lead to behaviors such as binge eating, finally resulting in excess caloric intake (Barson, 2020).

To provide a connection between food-seeking behavior and uncertainty, evidence on the effects of increasing such uncertainty on the proximal effect of food-seeking behavior, that is, food intake is necessary. In that regard, it was pointed out that, possibly because of survival mechanisms, environment uncertainty increased food intake and reduced energetic spending. Then, the sufficient functions to support such findings were discussed, emphasizing related structures and functions associated with each one. Obesity was associated with sharp delayed-discounting and ACC atrophy, which points towards a sub-optimal pairing between reward value assignments, given environment uncertainty levels. Also, the OH system role in foraging was discussed as a proximal cause of overfeeding. Together, this suggests that food-seeking behavior evolved to provide optimal decision-making strategies in uncertain and scarce environments. However, (1) when environment energetic density is high, such strategies would result in overfeeding, and (2) obesity in itself can impair homeostatic regulation by altering structures related to uncertainty and reward value processing. Previous points predict that underlying foraging mechanisms, in certain environments, can lead to obesity.

Models explaining food intake in obesity

Reinforcement learning models

Temporal-difference learning models state how agents can estimate reward values in uncertain environments. At each time-step, the agent computes the value of a given state considering: (1) the estimated value (randomly initiated at first), and (2) the temporal-difference error, which represents the distance between the estimate of state value and the actual reward obtained in such state.

$$V(S_t) \leftarrow V(S_t) + \alpha(Temporal\,Difference\,Error) \tag{1}$$

$V(S_t)$ denotes the estimated value at a given state, and α is used to model the agent learning rate. Additional parameter ρ has been proposed to model sensitivity to reward (Huys et al., 2013; Kroemer & Small, 2016), such that the temporal difference error accounts for the subjective value of obtained rewards.

$$Temporal\,Difference\,Error = \rho \times Reward - V(S_t) \tag{2}$$

Obese subjects had shown reduced dorsal striatum activity to food rewards, which has been interpreted as reduced pleasure for food. However, simulations under the previously presented model show another option. That is, obese subjects show heightened reward sensitivity but decreased learning rates, ending in a lowered state value estimation (Kroemer & Small, 2016). Modeled learning rates measures had shown that this is the case in obese subjects. Moreover, it points that negative prediction errors (the equivalent of temporal difference error) were used to a lesser extent than lean subjects, whereas positive errors showed no differences (Mathar et al., 2017). This can be interpreted as a difficulty to update reward or state values when the estimated reward is higher than the actual reward, possibly reflecting a short-term reward estimation.

It should be noted that more recent neuroimaging evidence points in favor of a hyper-reactivity of rewards circuitry, instead of hypo-reactivity. However, conclusions obtained by the model still hold, as such, hyper-reactivity is accompanied by a bias towards immediate rewards (Stice & Burger, 2019). In line with the reinforcement learning model presented, evidence from probabilistic learning paradigms in obese subjects shows a decreased impact of negatively valued choices on consequent behavioral adaptation (Kube et al., 2018). These seemingly opposing results can stem from, previously not considered, quadratic associations between BMI (body mass index) and reward sensitivity, where an inverted U-shape is observed as BMI increases (Horstmann et al., 2015). Taken together, this finding suggests that obesity overfeeding is not only reliant on increased reward sensitivity (more reward sensitivity is assumed to increase intake), but other parameters such as learning rates can determine the overall valuation of the reward, biasing decision-making to immediate rewards, that paired with highly palatable food can lead to excess caloric intake. This, because, while palatable food definition is not standardized (Fazzino et al., 2019), it can be assumed that they, typically, consist of high caloric density. However, there might be additional effects of palatable foods in decreasing taste sensitivity related brain areas, which in turn, might favor further intake (Yokum & Stice, 2019).

More complex models can include reward sensitivity in addition to different palatability indices of food encounters, effectively modeling the course of an agent with reward heterogeneity. Additionally, agents in this model are allowed to learn the value of different rewards values through different environments, such as highly-palatability, low-palatability, and mixed. Later on, the effects of the starting environment can be assessed. Results

suggest that starting in an abundance of highly-palatable food slows the learning of food reward values in the following environments (Hammond et al., 2012). While this model does not inform about the effect on weight or intake levels, it shows how agents react to initial conditions or, more generally, to non-stationary environments.

Delayed discounting models

Although the factors determining obesity as an outcome are multiple (Ang et al., 2013), it is reasonable to assume that the more immediate cause is excess intake relative to energetic demands. Moreover, excess intake is determined in an instance to instance basis, where a decision considering short and long-term benefits/risks must be made. With this in consideration, one can assume that obesity, in part, is caused by sub-optimal short/long-term benefit/risk assessments when making feeding decisions. If this was the case, as previously noted, areas that are related to computing options value in the short/long term, such as the ACC, should be in some way impaired.

Delayed discounting refers to the depreciation of a certain reward as a function of the time required to obtain it (da Matta et al., 2012). As such, it provides measures of how reward-related systems bias decision to the short or long term. Obese subjects show a robust tendency to steeply discount future rewards (Amlung et al., 2016), thus, favoring short-term rewards.

Furthermore, ACC, among other structures, shows relative atrophy in obese subjects (Raji et al., 2009; Wang et al., 2017), suggesting an impairment of the previously mentioned functions. These findings can be interpreted as if impairment in environment uncertainty assessment results in a preference for short-term rewards. If this were the case, palatable food sensory cues, which trigger food-intake, would dominate over more long-term modulated decisions, such as healthy food intake (Higgs, 2016).

Higher future rewards discounting paired with increased motivation to work for food, predict higher caloric intake (Rollins et al., 2010), and this effect seems to hold even for low energy-density food (Epstein et al., 2014). The rate of reward discounting, thus, informs about the predisposition to increased energetic intake, independent of possible food-property related effects. Similar effects have been found in children (Best et al., 2012), but not in adult males (Smulders et al., 2019). Moreover, these effects seem to be directly related to body fat (Rasmussen et al., 2010).

From uncertain environments to the cafeteria diet

(Mascia et al., 2019) (Hammond et al., 2012)

Conclusion

Figures

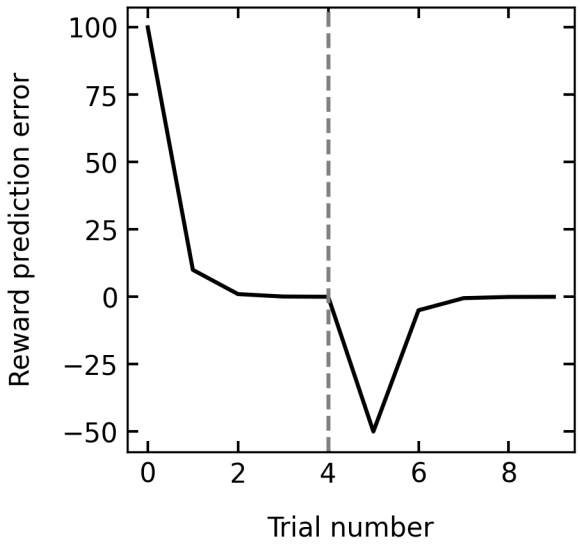

Figure 1: Figure shows a series of 10 trials, where from trial 0-4 the true reward value is 100, and for the remaining trial its 50. A very basic agent was simulated to update its estimates based on the reward prediction error. Initial estimates were set at 0. Notice how during trial 0-4 reward prediction errors are positive and decrease to 0, because the reward obtained was, initially, greater than the estimate, whereas, in trial 5, when reward changes to 50, the error becomes negative because the estimate was near 100

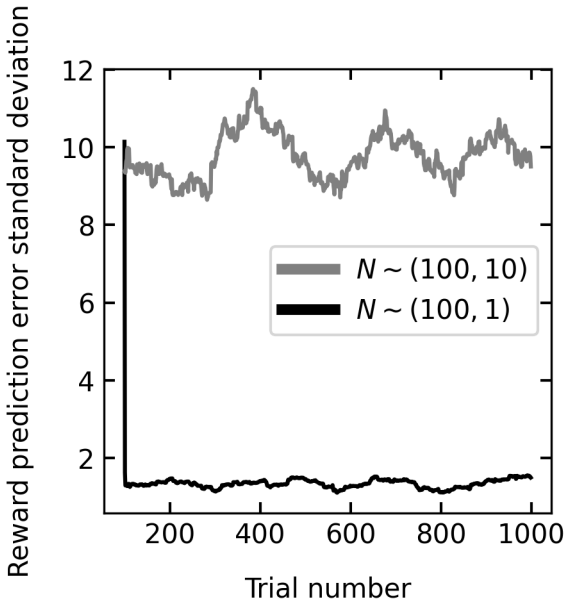

Figure 2: Simulated agent learning under two environments, (1) in gray, rewards are sampled from a normal distribution with mean = 100 and standard deviation = 10, (2) in black, mean = 100 and standard deviation = 1. Black and gray lines represent the reward prediction error rolling standard deviation over 100 trials. Notice how, over the trials, this 'signal' approximates to the underlying uncertainty of the distribution (using the standard deviation as measure of uncertainty).

References

Aitta-aho, T., Pappa, E., Burdakov, D., & Apergis-Schoute, J. (2016). Cellular activation of hypothalamic hypocretin/orexin neurons facilitates short-term spatial memory in mice. *Neurobiology of Learning and Memory*, *136*, 183–188. https://doi.org/10.1016/j.nlm.2016.10.005

Amlung, M., Petker, T., Jackson, J., Balodis, I., & MacKillop, J. (2016). Steep discounting of delayed monetary and food rewards in obesity: a meta-analysis. *Psychological Medicine*, *46*(11), 2423–2434. https://doi.org/10.1017/S0033291716000866

Ang, Y. N., Wee, B. S., Poh, B. K., & Ismail, M. N. (2013). Multifactorial Influences of Childhood Obesity. *Current Obesity Reports*, *2*(1), 10–22. https://doi.org/10.1007/s13679-012-0042-7

Anselme, P., & Güntürkün, O. (2019). How foraging works: Uncertainty magnifies food-seeking motivation. *Behavioral and Brain Sciences*, *42*, e35. https://doi.org/10.1017/S0140525X18000948

Anselme, P., Robinson, M. J. F., & Berridge, K. C. (2013). Reward uncertainty enhances incentive salience attribution as sign-tracking. *Behavioural Brain Research*, *238*, 53–61. https://doi.org/10.1016/j.bbr.2012.10.006

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, *47*(2/3), 235–256. https://doi.org/10.1023/A:1013689704352

Balleine, B. W., Delgado, M. R., & Hikosaka, O. (2007). The Role of the Dorsal Striatum in Reward and Decision-Making. *Journal of Neuroscience*, *27*(31), 8161–8165. https://doi.org/10.1523/JNEUROSCI.1554-07.2007

Barson, J. R. (2020). Orexin/hypocretin and dysregulated eating: Promotion of foraging behavior. *Brain Research*, *1731*, 145915. https://doi.org/10.1016/j.brainres.2018.08.018

Bayer, H. M., & Glimcher, P. W. (2005). Midbrain Dopamine Neurons Encode a Quantitative Reward Prediction Error Signal. *Neuron*, *47*(1), 129–141. https://doi.org/10.1016/j.neuron.2005.05.020

Bednekoff, P. A., & Houston, A. I. (1994). Avian daily foraging patterns: Effects of digestive constraints and variability. *Evolutionary Ecology*, *8*(1), 36–52. https://doi.org/10.1007/BF01237664

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), 1214–1221. https://doi.org/10.1038/nn1954

Best, J. R., Theim, K. R., Gredysa, D. M., Stein, R. I., Welch, R. R., Saelens, B. E., Perri, M. G., Schechtman, K. B., Epstein, L. H., & Wilfley, D. E. (2012). Behavioral economic predictors of overweight children's weight loss. *Journal of Consulting and Clinical Psychology*, *80*(6), 1086–1096. https://doi.org/10.1037/a0029827

Blanco, N. J., & Sloutsky, V. (2019). *Systematic Exploration and Uncertainty Dominate Young Children's Choices*. PsyArXiv. https://osf.io/72sfx

Chakroun, K., Mathar, D., Wiehler, A., Ganzer, F., & Peters, J. (2020). Dopaminergic modulation of the exploration/exploitation trade-off in human decision-making. *eLife*, *9*, e51260. https://doi.org/10.7554/eLife.51260

Charnov, E. L. (1976). Optimal foraging, the marginal value theorem. *Theoretical Population Biology*, *9*(2), 129–136. https://doi.org/10.1016/0040-5809(76)90040-X

Collins, L., Young, D. B., Davies, K., & Pearce, J. M. (1983). The Influence of Partial Reinforcement on Serial Autoshaping with Pigeons. *The Quarterly Journal of Experimental Psychology Section B*, *35*(4b), 275–290. https://doi.org/10.1080/14640748308400893

Craft, B. B. (2016). Risk-sensitive foraging: changes in choice due to reward quality and delay. *Animal Behaviour*, *111*, 41–47. https://doi.org/10.1016/j.anbehav.2015.09.030

Cuthill, I. C. (2000). Body mass regulation in response to changes in feeding predictability and overnight energy expenditure. *Behavioral Ecology*, *11*(2), 189–195. https://doi.org/10.1093/beheco/11.2.189

da Matta, A., Gonçalves, F. L., & Bizarro, L. (2012). Delay discounting: Concepts and measures. *Psychology & Neuroscience*, *5*(2), 135–146. https://doi.org/10.3922/j.psns.2012.2.03

Davis, J. F., Choi, D. L., & Benoit, S. C. (2010). Insulin, leptin and reward. *Trends in Endocrinology & Metabolism*, *21*(2), 68–74. https://doi.org/10.1016/j.tem.2009.08.004

Daw, N. D., Kakade, S., & Dayan, P. (2002). Opponent interactions between serotonin and dopamine. *Neural Networks*, *15*(4-6), 603–616. https://doi.org/10.1016/S0893-6080(02)00052-7

de Berker, A. O., Rutledge, R. B., Mathys, C., Marshall, L., Cross, G. F., Dolan, R. J., & Bestmann, S. (2016). Computations of uncertainty mediate acute stress responses in humans. *Nature Communications*, *7*(1), 10996. https://doi.org/10.1038/ncomms10996

De Groot, K., & Thurik, R. (2018). Disentangling Risk and Uncertainty: When Risk-Taking Measures Are Not About Risk. *Frontiers in Psychology*, *9*, 2194. https://doi.org/10.3389/fpsyg.2018.02194

Diederen, K. M. J., & Schultz, W. (2015). Scaling prediction errors to reward variability benefits error-driven learning in humans. *Journal of Neurophysiology*, *114*(3), 1628–

1640. https://doi.org/10.1152/jn.00483.2015

Enomoto, K., Matsumoto, N., Nakai, S., Satoh, T., Sato, T. K., Ueda, Y., Inokawa, H., Haruno, M., & Kimura, M. (2011). Dopamine neurons learn to encode the long-term value of multiple future rewards. *Proceedings of the National Academy of Sciences*, *108*(37), 15462–15467. https://doi.org/10.1073/pnas.1014457108

Epstein, L. H., Jankowiak, N., Fletcher, K. D., Carr, K. A., Nederkoorn, C., Raynor, H. A., & Finkelstein, E. (2014). Women who are motivated to eat and discount the future are more obese: BMI and Reinforcement Pathology. *Obesity*, *22*(6), 1394–1399. https://doi.org/10.1002/oby.20661

Even-Dar, E., & Mansour, Y. (2001). Learning Rates for Q-Learning. In D. Helmbold & B. Williamson (Eds.), *Computational Learning Theory* (Vol. 2111, pp. 589–604). Springer Berlin Heidelberg. http://link.springer.com/10.1007/3-540-44581-1_39

Faraji, M., Preuschoff, K., & Gerstner, W. (2017). Balancing New Against Old Information: The Role of Surprise in Learning. *arXiv:1606.05642 [Cs, Q-Bio, Stat]*. http://arxiv.org/abs/1606.05642

Fazzino, T. L., Rohde, K., & Sullivan, D. K. (2019). Hyper-Palatable Foods: Development of a Quantitative Definition and Application to the US Food System Database. *Obesity*, *27*(11), 1761–1768. https://doi.org/10.1002/oby.22639

Filipowicz, A. L., Glaze, C. M., Kable, J. W., & Gold, J. I. (2020). Pupil diameter encodes the idiosyncratic, cognitive complexity of belief updating. *eLife*, *9*, e57872. https://doi.org/10.7554/eLife.57872

Fiorillo, C. D. (2003). Discrete Coding of Reward Probability and Uncertainty by Dopamine Neurons. *Science*, *299*(5614), 1898–1902. https://doi.org/10.1126/science.1077349

Flagel, S. B. (2014). Sign-Tracking. In I. P. Stolerman & L. H. Price (Eds.), *Encyclopedia of Psychopharmacology* (pp. 1–7). Springer Berlin Heidelberg. http://link.springer.com/10.1007/978-3-642-27772-6_7020-1

Fokidis, H. B., des Roziers, M. B., Sparr, R., Rogowski, C., Sweazea, K., & Deviche, P. (2012). Unpredictable food availability induces metabolic and hormonal changes independent of food intake in a sedentary songbird. *Journal of Experimental Biology*, *215*(16), 2920–2930. https://doi.org/10.1242/jeb.071043

Forkman, B. A. (1993). The Effect of Uncertainty On the Food Intake of the Mongolian Gerbil. *Behaviour*, *124*(3-4), 197–206. https://doi.org/10.1163/156853993X00579

Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301. https://doi.org/10.1016/j.tics.2009.04.005

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42. https://doi.org/10.1016/j.cognition.2017.12.014

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, *108*(Supplement_3), 15647–15654. https://doi.org/10.1073/pnas.1014269108

Hafenbrädl, S., Waeger, D., Marewski, J. N., & Gigerenzer, G. (2016). Applied Decision Making With Fast-and-Frugal Heuristics. *Journal of Applied Research in Memory and Cognition*, *5*(2), 215–231. https://doi.org/10.1016/j.jarmac.2016.04.011

Hammond, R. A., Ornstein, J. T., Fellows, L. K., Dubé, L., Levitan, R., & Dagher, A. (2012). A model of food reward learning with dynamic reward exposure. *Frontiers in Computational Neuroscience*, *6*. https://doi.org/10.3389/fncom.2012.00082

Harris, T. R., Chapman, C. A., & Monfort, S. L. (2010). Small folivorous primate groups exhibit behavioral and physiological effects of food scarcity. *Behavioral Ecology*, *21*(1), 46–56. https://doi.org/10.1093/beheco/arp150

Haruno, M., & Kawato, M. (2006). Different neural correlates of reward expectation and reward expectation error in the putamen and caudate nucleus during stimulus-action-reward association learning. *Journal of Neurophysiology*, *95*(2), 948–959. https://doi.org/10.1152/jn.00382.2005

Higgs, S. (2016). Cognitive processing of food rewards. *Appetite*, *104*, 10–17. https://doi.org/10.1016/j.appet.2015.10.003

Hills, T. T., Kalff, C., & Wiener, J. M. (2013). Adaptive Lévy Processes and Area-Restricted Search in Human Foraging. *PLoS ONE*, *8*(4), e60488. https://doi.org/10.1371/journal.pone.0060488

Horstmann, A., Fenske, W. K., & Hankir, M. K. (2015). Argument for a non-linear relationship between severity of human obesity and dopaminergic tone: Relationship between obesity and dopaminergic tone. *Obesity Reviews*, *16*(10), 821–830. https://doi.org/10.1111/obr.12303

Huettel, S. A. (2005). Decisions under Uncertainty: Probabilistic Context Influences Activation of Prefrontal and Parietal Cortices. *Journal of Neuroscience*, *25*(13), 3304–3311. https://doi.org/10.1523/JNEUROSCI.5070-04.2005

Humphries, N. E., & Sims, D. W. (2014). Optimal foraging strategies: Lévy walks balance searching and patch exploitation under a very broad range of conditions. *Journal of Theoretical Biology*, *358*, 179–193. https://doi.org/10.1016/j.jtbi.2014.05.032

Huys, Q. J., Pizzagalli, D. A., Bogdan, R., & Dayan, P. (2013). Mapping anhedonia onto reinforcement learning: a behavioural meta-analysis. *Biology of Mood & Anxiety Disorders*, *3*(1), 12. https://doi.org/10.1186/2045-5380-3-12

Kroemer, N. B., & Small, D. M. (2016). Fuel not fun: Reinterpreting attenuated brain responses to reward in obesity. *Physiology & Behavior*, *162*, 37–45. https://doi.org/10.1016/j.physbeh.2016.04.020

Kube, J., Mathar, D., Horstmann, A., Kotz, S. A., Villringer, A., & Neumann, J. (2018). Altered monetary loss processing and reinforcement-based learning in individuals with obesity. *Brain Imaging and Behavior*, *12*(5), 1431–1449. https://doi.org/10.1007/s11682-017-9786-8

Macleod, R., Barnett, P., Clark, J. A., & Cresswell, W. (2005). Body mass change strategies in blackbirds Turdus merula: the starvation-predation risk trade-off. *Journal of Animal Ecology*, *74*(2), 292–302. https://doi.org/10.1111/j.1365-2656.2005.00923.x

Mascia, P., Neugebauer, N. M., Brown, J., Bubula, N., Nesbitt, K. M., Kennedy, R. T., & Vezina, P. (2019). Exposure to conditions of uncertainty promotes the pursuit of amphetamine. *Neuropsychopharmacology*, *44*(2), 274–280. https://doi.org/10.1038/s41386-018-0099-4

Mathar, D., Neumann, J., Villringer, A., & Horstmann, A. (2017). Failing to learn from negative prediction errors: Obesity is associated with alterations in a fundamental neural learning mechanism. *Cortex*, *95*, 222–237. https://doi.org/10.1016/j.cortex.2017.08.022

Mikhael, J. G., & Bogacz, R. (2016). Learning Reward Uncertainty in the Basal Ganglia. *PLOS Computational Biology*, *12*(9), e1005062. https://doi.org/10.1371/journal.pcbi.1005062

Moiron, M., Mathot, K. J., & Dingemanse, N. J. (2018). To eat and not be eaten: diurnal mass gain and foraging strategies in wintering great tits. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1874), 20172868. https://doi.org/10.1098/rspb.2017.2868

Muranishi, M., Inokawa, H., Yamada, H., Ueda, Y., Matsumoto, N., Nakagawa, M., & Kimura, M. (2011). Inactivation of the putamen selectively impairs reward history-based action selection. *Experimental Brain Research*, *209*(2), 235–246. https://doi.org/10.1007/s00221-011-2545-y

Namdari, A., & Li, Z. (. (2019). A review of entropy measures for uncertainty quantification of stochastic processes. *Advances in Mechanical Engineering*, *11*(6), 168781401985735. https://doi.org/10.1177/1687814019857350

Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., & Gold, J. I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature Neuroscience*, *15*(7), 1040–1046. https://doi.org/10.1038/nn.3130

Nonomura, S., Nishizawa, K., Sakai, Y., Kawaguchi, Y., Kato, S., Uchigashima, M., Watanabe, M., Yamanaka, K., Enomoto, K., Chiken, S., Sano, H., Soma, S., Yoshida, J., Samejima, K., Ogawa, M., Kobayashi, K., Nambu, A., Isomura, Y., & Kimura, M. (2018). Monitoring and Updating of Action Selection for Goal-Directed Behavior through the Striatal Direct and Indirect Pathways. *Neuron*, *99*(6), 1302–1314.e5. https://doi.org/10.1016/j.neuron.2018.08.002

Payzan-LeNestour, E., Dunne, S., Bossaerts, P., & O'Doherty, J. (2013). The Neural Representation of Unexpected Uncertainty during Value-Based Decision Making. *Neuron*, *79*(1), 191–201. https://doi.org/10.1016/j.neuron.2013.04.037

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*(6), 532–552. https://doi.org/10.1037/0033-295X.87.6.532

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*(7106), 1042–1045. https://doi.org/10.1038/nature05051

Polo, V. (2002). Daily body mass regulation in dominance-structured coal tit (Parus ater) flocks in response to variable food access: a laboratory study. *Behavioral Ecology*, *13*(5), 696–704. https://doi.org/10.1093/beheco/13.5.696

Preuschoff, K., & Bossaerts, P. (2007). Adding Prediction Risk to the Theory of Reward Learning. *Annals of the New York Academy of Sciences*, *1104*(1), 135–146. https://doi.org/10.1196/annals.1390.005

Preuschoff, K., Bossaerts, P., & Quartz, S. R. (2006). Neural Differentiation of Expected Reward and Risk in Human Subcortical Structures. *Neuron*, *51*(3), 381–390. https://doi.org/10.1016/j.neuron.2006.06.024

Raji, C. A., Ho, A. J., Parikshak, N. N., Becker, J. T., Lopez, O. L., Kuller, L. H., Hua, X., Leow, A. D., Toga, A. W., & Thompson, P. M. (2009). Brain structure and obesity. *Human Brain Mapping*, NA–NA. https://doi.org/10.1002/hbm.20870

Rasmussen, E. B., Lawyer, S. R., & Reilly, W. (2010). Percent body fat is related to delay and probability discounting for food in humans. *Behavioural Processes*, *83*(1), 23–30. https://doi.org/10.1016/j.beproc.2009.09.001

Robinson, M. J. F., Anselme, P., Fischer, A. M., & Berridge, K. C. (2014). Initial uncertainty in Pavlovian reward prediction persistently elevates incentive salience and extends sign-tracking to normally unattractive cues. *Behavioural Brain Research*, *266*, 119–130. https://doi.org/10.1016/j.bbr.2014.03.004

Robinson, M. J. F., Anselme, P., Suchomel, K., & Berridge, K. C. (2015). Amphetamine-induced sensitization and reward uncertainty similarly enhance incentive salience for

conditioned cues. *Behavioral Neuroscience*, *129*(4), 502–511. https://doi.org/10.1037/bne0000064

Rollins, B. Y., Dearing, K. K., & Epstein, L. H. (2010). Delay discounting moderates the effect of food reinforcement on energy intake among non-obese women. *Appetite*, *55*(3), 420–425. https://doi.org/10.1016/j.appet.2010.07.014

Rushworth, M. F. S., & Behrens, T. E. J. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. *Nature Neuroscience*, *11*(4), 389–397. https://doi.org/10.1038/nn2066

Samejima, K. (2005). Representation of Action-Specific Reward Values in the Striatum. *Science*, *310*(5752), 1337–1340. https://doi.org/10.1126/science.1115270

Schultz, W. (2016). Dopamine reward prediction error coding. *Dialogues in Clinical Neuroscience*, *18*(1), 23–32. https://doi.org/27069377

Schultz, W., Dayan, P., & Montague, P. R. (1997). A Neural Substrate of Prediction and Reward. *Science*, *275*(5306), 1593–1599. https://doi.org/10.1126/science.275.5306.1593

Schultz, W., Preuschoff, K., Camerer, C., Hsu, M., Fiorillo, C. D., Tobler, P. N., & Bossaerts, P. (2008). Explicit neural signals reflecting reward uncertainty. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1511), 3801–3811. https://doi.org/10.1098/rstb.2008.0152

Shenhav, A., Cohen, J. D., & Botvinick, M. M. (2016). Dorsal anterior cingulate cortex and the value of control. *Nature Neuroscience*, *19*(10), 1286–1291. https://doi.org/10.1038/nn.4384

Silva, F. J., Silva, K., & Pear, J. J. (1992). SIGN- VERSUS GOAL-TRACKING: EFFECTS OF CONDITIONED-STIMULUS-TO-UNCONDITIONED-STIMULUS DISTANCE. *Journal of the Experimental Analysis of Behavior*, *57*(1), 17–31. https://doi.org/10.1901/jeab.1992.57-17

Smulders, T. V., Boswell, T., & Henderson, L. J. (2019). "How Foraging Works": Let's not forget the physiological mechanisms of energy balance. *Behavioral and Brain Sciences*, *42*, e51. https://doi.org/10.1017/S0140525X1800198X

Soltani, A., & Izquierdo, A. (2019). Adaptive learning under expected and unexpected uncertainty. *Nature Reviews Neuroscience*, *20*(10), 635–644. https://doi.org/10.1038/s41583-019-0180-y

Somerville, L. H., Sasse, S. F., Garrad, M. C., Drysdale, A. T., Abi Akar, N., Insel, C., & Wilson, R. C. (2017). Charting the expansion of strategic exploratory behavior during adolescence. *Journal of Experimental Psychology: General*, *146*(2), 155–164. https://doi.org/10.1037/xge0000250

Stagner, J. P., & Zentall, T. R. (2010). Suboptimal choice behavior by pigeons. *Psychonomic Bulletin & Review*, *17*(3), 412–416. https://doi.org/10.3758/PBR.17.3.412

Stice, E., & Burger, K. (2019). Neural vulnerability factors for obesity. *Clinical Psychology Review*, *68*, 38–53. https://doi.org/10.1016/j.cpr.2018.12.002

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: an introduction* (Second edition). The MIT Press.

Taha, S. A. (2010). Preference or fat? Revisiting opioid effects on food intake. *Physiology & Behavior*, *100*(5), 429–437. https://doi.org/10.1016/j.physbeh.2010.02.027

Tajima, S., Drugowitsch, J., & Pouget, A. (2016). Optimal policy for value-based decision-making. *Nature Communications*, *7*(1), 12400. https://doi.org/10.1038/ncomms12400

Takahashi, Y. K., Batchelor, H. M., Liu, B., Khanna, A., Morales, M., & Schoenbaum, G. (2017). Dopamine Neurons Respond to Errors in the Prediction of Sensory Features of Expected Rewards. *Neuron*, *95*(6), 1395–1405.e3. https://doi.org/10.1016/j.neuron.2017.08.025

Tanaka, S. C., Samejima, K., Okada, G., Ueda, K., Okamoto, Y., Yamawaki, S., & Doya, K. (2006). Brain mechanism of reward prediction under predictable and unpredictable environmental dynamics. *Neural Networks*, *19*(8), 1233–1241. https://doi.org/10.1016/j.neunet.2006.05.039

Toshinai, K., Date, Y., Murakami, N., Shimada, M., Mondal, M. S., Shimbara, T., Guan, J.-L., Wang, Q.-P., Funahashi, H., Sakurai, T., Shioda, S., Matsukura, S., Kangawa, K., & Nakazato, M. (2003). Ghrelin-Induced Food Intake Is Mediated via the Orexin Pathway. *Endocrinology*, *144*(4), 1506–1512. https://doi.org/10.1210/en.2002-220788

van den Berg, P., & Wenseleers, T. (2018). Uncertainty about social interactions leads to the evolution of social heuristics. *Nature Communications*, *9*(1), 2151. https://doi.org/10.1038/s41467-018-04493-1

Wajnberg, E., Fauvergue, X., & Pons, O. (2000). Patch leaving decision rules and the Marginal Value Theorem: an experimental analysis and a simulation model. *Behavioral Ecology*, *11*(6), 577–586. https://doi.org/10.1093/beheco/11.6.577

Wake, S. J., & Izuma, K. (2017). A common neural code for social and monetary rewards in the human striatum. *Social Cognitive and Affective Neuroscience*, *12*(10), 1558–1564. https://doi.org/10.1093/scan/nsx092

Wang, H., Wen, B., Cheng, J., & Li, H. (2017). Brain Structural Differences between Normal and Obese Adults and their Links with Lack of Perseverance, Negative Urgency, and Sensation Seeking. *Scientific Reports*, *7*(1), 40595. https://doi.org/10.1038/srep40595

Watabe-Uchida, M., Eshel, N., & Uchida, N. (2017). Neural Circuitry of Reward Prediction Error. *Annual Review of Neuroscience*, *40*(1), 373–394. https://doi.org/10.1146/annurev-neuro-072116-031109

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore–exploit dilemma. *Journal of Experimental Psychology: General*, *143*(6), 2074–2081. https://doi.org/10.1037/a0038199

Wolf, G. (2009). Orexins: A Newly Discovered Family of Hypothalamic Regulators of Food Intake. *Nutrition Reviews*, *56*(6), 172–173. https://doi.org/10.1111/j.1753-4887.1998.tb06131.x

Wosniack, M. E., Santos, M. C., Raposo, E. P., Viswanathan, G. M., & da Luz, M. G. E. (2017). The evolutionary origins of Lévy walk foraging. *PLOS Computational Biology*, *13*(10), e1005774. https://doi.org/10.1371/journal.pcbi.1005774

Yokum, S., & Stice, E. (2019). Weight gain is associated with changes in neural response to palatable food tastes varying in sugar and fat and palatable food images: a repeated-measures fMRI study. *The American Journal of Clinical Nutrition*, *110*(6), 1275–1286. https://doi.org/10.1093/ajcn/nqz204

Yu, A. J., & Dayan, P. (2005). Uncertainty, Neuromodulation, and Attention. *Neuron*, *46*(4), 681–692. https://doi.org/10.1016/j.neuron.2005.04.026