Rewards are the most crucial objects for life. Their function is to make us eat, drink, and mate. Species with brains that allow them to get better rewards will win in evolution. This is what our brain does, acquire rewards, and do it in the best possible way. It may well be the reason why brains have evolved. Brains allow multicellular organisms to move about the world. By displacing themselves they can access more rewards than happen to come along by chance, thus enhancing their chance of survival and reproduction. However, movement alone does not get them any food or mating partners. It is necessary to identify stimuli, objects, events, situations, and activities that lead to the best nutrients and mating partners. Brains make individuals learn, select, approach, and consume the best rewards for survival and reproduction and thus make them succeed in evolutionary selection. To do so, the brain needs to identify the reward value of objects for survival and reproduction, and then direct the acquisition of these reward objects through learning, approach, choices, and positive emotions. Sensory discrimination and control of movements serve this prime role of the brain. For these functions, nature has endowed us with explicit neuronal reward signals that process all crucial aspects of reward functions.

Rewards are not defined by their physical properties but by the behavioral reactions they induce. Therefore, we need behavioral theories that provide concepts of reward functions. The theoretical concepts can be used for making testable hypotheses for experiments and for interpreting the results. Thus the field of reward and decision-making is not only hypothesis driven but also concept driven. The field of reward and decision-making benefits from well-developed theories of behavior as the study of sensory systems benefits from signal detection theory and the study of the motor system benefits from an understanding of mechanics. Reward theories are particularly important because of the absence of specific sensory receptors for reward, which would have provided basic physical definitions. Thus the theories help to overcome the limited explanatory power of physical reward parameters and emphasize the requirement for behavioral assessment of the reward parameters studied. These theories make disparate data consistent and coherent and thus help to avoid seemingly intuitive but paradoxical explanations.

Theories of reward function employ a few basic, fundamental variables such as subjective reward value derived from measurable behavior. This variable condenses all crucial factors of reward function and allows quantitative formal-
ization that characterizes and predicts a large variety of behavior. Importantly, this variable is hypothetical and does not exist in the external physical world. However, it is implemented in the brain in various neuronal reward signals, and thus does seem to have a physical basis. Although sophisticated forms of reward and decision processes are far more fascinating than arcane fundamental variables, their investigation may be crucial for understanding reward processing. Where would we be without the discovery of the esoteric electron by J. J. Thompson 1897 in the Cambridge Cavendish Laboratory? Without this discovery, the microprocessor and the whole internet would be impossible. Or, if we did not know about electromagnetic waves, we might assume a newsreader sitting inside the radio while sipping our morning coffee. This review is particularly concerned with fundamental reward variables, first concerning learning and then related to decision-making.

The reviewed work concerns primarily neurophysiological studies on single neurons in monkeys whose sophisticated behavioral repertoire allows well detailed, quantitative behavioral assessments while controlling confounds from sensory processing, movements, and attention. Thus I am approaching reward processing from the point of view of the tip of a microelectrode, one neuron at a time, thousands of them over the years, in rhesus’ brains with more than two billion neurons. I apologize to the authors whose work I have not been able to cite in full, as there is a large number of recent studies on the subject and I am selecting these studies by their contribution to the concepts being treated here.

II. REWARD FUNCTIONS

A. Proximal Reward Functions Are Defined by Behavior

We have sensory receptors that react to environmental events. The retina captures electromagnetic waves in a limited range. Optical physics, physical chemistry, and biochemistry help us to understand how the waves enter the eye, how the photons affect the ion channels in the retinal photoreceptors, and how the ganglion cells transmit the visual message to the brain. Thus sensory receptors define the functions of the visual system by translating the energy from environmental events into action potentials and sending them to the brain. The same holds for touch, pain, hearing, smell, and taste. If there are no receptors for particular environmental energies, we do not sense them. Humans do not feel magnetic fields, although some fish do. Thus physics and chemistry are a great help for defining and investigating the functions of sensory systems.

Rewards have none of that. Take rewarding stimuli and objects: we see them, feel them, taste them, smell them, or hear them. They affect our body through all sensory systems, but there is not a specific receptor that would capture the particular motivational properties of rewards. As reward functions cannot be explained by object properties alone, physics and chemistry are only of limited help, and we cannot investigate reward processing by looking at the properties of reward receptors. Instead, rewards are defined by the particular behavioral reactions they induce. Thus, to understand reward function, we need to study behavior. Behavior becomes the key tool for investigating reward function, just as a radio telescope is a key tool for astronomy.

The word reward has almost mystical connotations and is the subject of many philosophical treatises, from the ethics of the utilitarian philosophy of Jeremy Bentham (whose embalmed body is displayed in University College London) and John Stuart Mill to the contemporary philosophy of science of Tim Schroeder (39, 363, 514). More commonly, the man on the street views reward as a bonus for exceptional performance, like chocolate for a child getting good school marks, or as something that makes us happy. These descriptions are neither complete nor practical for scientific investigations. The field has settled on a number of well-defined reward functions that have allowed an amazing advance in knowledge on reward processing and have extended these investigations into economic decision-making. We are dealing with three, closely interwoven, functions of reward, namely, learning, approach behavior and decision-making, and pleasure.

1. Learning

Rewards have the potential to produce learning. Learning is Pavlov’s main reward function (423). His dog salivates to a bell when a sausage often follows, but it does not salivate just when a bell rings without consequences. The animal’s reaction to the initially neutral bell has changed because of the sausage. Now the bell predicts the sausage. No own action is required, as the sausage comes for free, and the learning happens also for free. Thus Pavlovian learning (classical conditioning) occurs automatically, without the subject’s own active participation, other than being awake and mildly attentive. Then there is Thorndike’s cat that runs around the cage and, among other things, presses a lever and suddenly gets some food (589). The food is great, and the cat presses again, and again, with increasing enthusiasm. The cat comes back for more. This is instrumental or operant learning. It requires an own action; otherwise, no reward will come and no learning will occur. Requiring an action is a major difference from Pavlovian learning. Thus operant learning is about actions, whereas Pavlovian learning is about stimuli. The two learning mechanisms can be distinguished schematically but occur frequently together and constitute the building blocks for behavioral reactions to rewards.
Rewards in operant conditioning are positive reinforcers. They increase and maintain the frequency and strength of the behavior that leads to them. The more reward Thorndike’s cat gets, the more it will press the lever. Reinforcers do not only strengthen and maintain behavior for the cat but also for obtaining stimuli, objects, events, activities, and situations as different as beer, whisky, alcohol, relaxation, beauty, mating, babies, social company, and hundreds of others. Operant behavior gives a good definition for rewards. Anything that makes an individual come back for more is a positive reinforcer and therefore a reward. Although it provides a good definition, positive reinforcement is only one of several reward functions.

2. Approach behavior and decision-making

Rewards are attractive. They are motivating and make us exert an effort. We want rewards; we do not usually remain neutral when we encounter them. Rewards induce approach behavior, also called appetitive or preparatory behavior, and consummatory behavior. We want to get closer when we encounter them, and we prepare to get them. We cannot get the meal, or a mating partner, if we do not approach them. Rewards usually do not come alone, and we often can choose between different rewards. We find some rewards more attractive than others and select the best reward. Thus we value rewards and then decide between them to get the best value. Then we consume them. So, rewards are attractive and elicit approach behavior that helps to consume the reward. Thus any stimulus, object, event, activity, or situation that has the potential to make us approach and consume it is by definition a reward.

3. Positive emotions

Rewards have the potential to elicit positive emotions. The foremost emotion evoked by rewards is pleasure. We enjoy having a good meal, watching an interesting movie, or meeting a lovely person. Pleasure constitutes a transient response that may lead to the longer lasting state of happiness. There are different degrees and forms of pleasure. Water is pleasant for a thirsty person, and food for a hungry one. The rewarding effects of taste are based on the pleasure it evokes. Winning in a big lottery is even more pleasant. But many enjoyments differ by more than a few degrees. The feeling of high that is experienced by sports people during running or swimming, the lust evoked by encountering a ready mating partner, a sexual orgasm, the euphoria reported by drug users, and the parental affection to babies constitute different forms (qualities) rather than degrees of pleasure (quantities).

Once we have experienced the pleasure from a reward, we may form a desire to obtain it again. When I am thirsty or hungry and know that water or food helps, I desire them. Different from such specific desire, there are also desires for imagined or even impossible rewards, such as flying to Mars, in which cases desires become wishes (514). Desire requires a prediction, or at least a representation, of reward and constitutes an active process that is intentional [in being about something (529)]. Desire makes behavior purposeful and directs it towards identifiable goals. Thus desire is the emotion that helps to actively direct behavior towards known rewards, whereas pleasure is the passive experience that derives from a received or anticipated reward. Desire has multiple relations to pleasure; it may be pleasant in itself (I feel a pleasant desire), and it may lead to pleasure (I desire to obtain a pleasant object). Thus pleasure and desire have distinctive characteristics but are closely intertwined. They constitute the most important positive emotions induced by rewards. They prioritize our conscious processing and thus constitute important components of behavioral control. These emotions are also called liking (for pleasure) and wanting (for desire) in addiction research (471) and strongly support the learning and approach generating functions of reward.

Despite their immense power in reward function, pleasure and desire are very difficult to assess in an objectively measurable manner, which is an even greater problem for scientific investigations on animals, despite attempts to anthropomorphize (44). We do not know exactly what other humans feel and desire, and we know even less what animals feel. We can infer pleasure from behavioral responses that are associated with verbal reports about pleasure in humans. We could measure blood pressure, heart rate, skin resistance, or pupil diameter as manifestations of pleasure or desire, but they occur with many different emotions and thus are unspecific. Some of the stimuli and events that are pleasurable in humans may not even evoke pleasure in animals but act instead through innate mechanisms. We simply do not know. Nevertheless, the invention of pleasure and desire by evolution had the huge advantage of allowing a large number of stimuli, objects, events, situations, and activities to be attractive. This mechanism importantly supports the primary reward functions in obtaining essential substances and mating partners.

4. Potential

Rewards have the potential to produce learning, approach, decisions, and positive emotions. They are rewards even if their functions are not evoked at a given moment. For example, operant learning occurs only if the subject makes the operant response, but the reward remains a reward even if the subject does not make the operant response and the reward cannot exert its learning function. Similarly an object that has the potential to induce approach or make me happy or desire it is a reward, without necessarily doing it every time because I am busy or have other reasons not to engage. Pavlovian conditioning of approach behavior, which occurs every time a reward is encountered as long as it evokes at least minimal attention, nicely shows this.
sory, gustatory, and olfactory) (basic sensory components (visual, auditory, somatosensory, gustatory, and olfactory) consist of several major components. First, rewards have a value component that determines the positively motivating effects of rewards and is not contained in, nor explained by, the sensory and attentional components (FIGURE 1, right). This component reflects behavioral preferences and thus is subjective and only partially determined by physical parameters. Only this component constitutes what we understand as a reward. It mediates the specific behavioral reinforcing, approach generating, and emotional effects of rewards that are crucial for the organism’s survival and reproduction, whereas all other components are only supportive of these functions.

5. Punishment

The second large category of motivating events besides rewards is punishment. Punishment produces negative Pavlovian learning and negative operant reinforcement, passive and active avoidance behavior and negative emotions like fear, disgust, sadness, and anger (143). Finer distinctions separate punishment (reduction of response strength, passive avoidance) from negative reinforcement (enhancing response strength, active avoidance).

6. Reward components

Rewarding stimuli, objects, events, situations, and activities consist of several major components. First, rewards have basic sensory components (visual, auditory, somatosensory, gustatory, and olfactory) (FIGURE 1, left), with physical parameters such as size, form, color, position, viscosity, acidity, and others. Food and liquid rewards contain chemical substances necessary for survival such as carbohydrates, proteins, fats, minerals, and vitamins, which contain physically measurable quantities of molecules. These sensory components act via specific sensory receptors on the brain. Some rewards consist of situations, which are detected by cognitive processes, or activities involving motor processes, which also constitute basic components analogous to sensory ones. Second, rewards are salient and thus elicit attention, which are manifested as orienting responses (FIGURE 1, middle). The salience of rewards derives from three principal factors, namely, their physical intensity and impact (physical salience), their novelty and surprise (novelty/surprise salience), and their general motivational impact shared with punishers (motivational salience). A separate form not included in this scheme, incentive salience, primarily addresses dopamine function in addiction and refers only to approach behavior (as opposed to learning) and thus to reward and not punishment (471). The term is at odds with current results on the role of dopamine in learning (see below) and reflects an earlier assumption of attentional dopamine function based on an initial phasic response component before distinct dopamine response components were recognized (see below). Third, rewards have a value component that determines the positively motivating effects of rewards and is not contained in, nor explained by, the sensory and attentional components (FIGURE 1, right). This component reflects behavioral preferences and thus is subjective and only partially determined by physical parameters. Only this component constitutes what we understand as a reward. It mediates the specific behavioral reinforcing, approach generating, and emotional effects of rewards that are crucial for the organism’s survival and reproduction, whereas all other components are only supportive of these functions.

The major reward components together ensure maximal reward acquisition. Without the sensory component, reward discrimination would be difficult; without the attentional components, reward processing would be insufficiently prioritized; and without valuation, useless objects would be pursued. In practical reward experiments, the value component should be recognized as a distinct variable in the design and distinguished and uncorrelated from the sensory and attentional components.

The reward components can be divided into external components that reflect the impact of environmental stimuli, objects and events on the organism, and internal components generated by brain function. The sensory components are external, as they derive from external events and allow stimulus identification before evaluation can begin. In analogy, the external physical salience components lead to stimulus-driven attention. The foremost internal component is reward value. It is not inherently attached to stimuli, objects, events, situations, and activities but reflects the brain’s assessment of their usefulness for survival and reproduction. Value cannot be properly defined by physical reward parameters but is represented in subjective preferences that are internal, private, unobservable, and incomparable between individuals. These preferences are elicited by approach behavior and choices that can be objectively measured. The internal nature of value extends to its associated motivational salience. Likewise, reward predictors are not
hardwired to outside events but require neuronal learning and memory processes, as does novelty/surprise salience which relies on comparisons with memorized events. Reward predictors generate top-down, cognitive attention that establishes a saliency map of the environment before the reward occurs. Further internal reward components are cognitive processes that identify potentially rewarding environmental situations, and motor processes mediating intrinsically rewarding movements.

B. Distal Reward Function Is Evolutionary Fitness

Modern biological theory conjectures that the currently existing organisms are the result of evolutionary competition. Advancing the idea about survival of the fittest organisms, Richard Dawkins stresses gene survival and propagation as the basic mechanism of life (114). Only genes that lead to the fittest phenotype will make it. The phenotype is selected based on behavior that maximizes gene propagation. To do so, the phenotype must survive and generate offspring, and be better at it than its competitors. Thus the ultimate, distal function of rewards is to increase evolutionary fitness by ensuring survival of the organism and reproduction. Then the behavioral reward functions of the present organisms are the result of evolutionary selection of phenotypes that maximize gene propagation. Learning, approach, economic decisions, and positive emotions are the proximal functions through which phenotypes obtain the necessary nutrients for survival, mating, and care for offspring.

Behavioral reward functions have evolved to help individuals to propagate their genes. Individuals need to live well and long enough to reproduce. They do so by ingesting the substances that make their bodies function properly. The substances are contained in solid and liquid forms, called foods and drinks. For this reason, foods and drinks are rewards. Additional rewards, including those used for economic exchanges, ensure sufficient food and drink supply. Mating and gene propagation is supported by powerful sexual attraction. Additional properties, like body form, enhance the chance to mate and nourish and defend offspring and are therefore rewards. Care for offspring until they can reproduce themselves helps gene propagation and is rewarding; otherwise, mating is useless. As any small edge will ultimately result in evolutionary advantage (112), additional reward mechanisms like novelty seeking and exploration widen the spectrum of available rewards and thus enhance the chance for survival, reproduction, and ultimate gene propagation. These functions may help us to obtain the benefits of distant rewards that are determined by our own interests and not immediately available in the environment. Thus the distal reward function in gene propagation and evolutionary fitness defines the proximal reward functions that we see in everyday behavior. That is why foods, drinks, mates, and offspring are rewarding.

The requirement for reward seeking has led to the evolution of genes that define brain structure and function. This is what the brain is made for: detecting, seeking, and learning about rewards in the environment by moving around, identifying stimuli, valuing them, and acquiring them through decisions and actions. The brain was not made for enjoying a great meal; it was made for getting the best food for survival, and one of the ways to do that is to make sure that people are attentive and appreciate what they are eating.

C. Types of Rewards

The term reward has many names. Psychologists call it positive reinforcer because it strengthens behaviors that lead to reward, or they call it outcome of behavior or goal of action. Economists call it a good or commodity and assess the subjective value for the decision maker as utility. We now like to identify the kinds of stimulus, object, event, activity, and situation that elicit the proximal functions of learning, approach, decision-making, and positive emotions and thus serve the ultimate, distal reward function of evolutionary fitness.

1. Primary homeostatic and reproductive rewards

To ensure gene propagation, the primary rewards mediate the survival of the individual gene carrier and her reproduction. These rewards are foods and liquids that contain the substances necessary for individual survival, and the activities necessary to mate, produce offspring, and care for the offspring. They are attractive and the main means to achieve evolutionary fitness in all animals and humans. Primary food and liquid rewards serve to correct homeostatic imbalances. They are the basis for Hull’s drive reduction theory (242) that, however, would not apply to rewards that are not defined by homeostasis. Sexual behavior follows hormonal imbalances, at least in men, but is also strongly based on pleasure. To acquire and follow these primary alimentary and mating rewards is the main reason why the brain’s reward system has evolved in the first place. Note that “primary” reward does not refer to the distinction between unconditioned versus conditioned reward; indeed, most primary rewards are learned and thus conditioned (foods are primary rewards that are typically learnt).

2. Nonprimary rewards

All other rewards serve to enhance the function of primary alimentary and mating rewards and thus enhance the chance for survival, reproduction, and evolutionary selection. Even though they are not homeostatic or reproductive rewards, they are rewards in their own rights. These nonprimary rewards can be physical, tangible objects like money, sleek cars, or expensive jewelry, or material liquids like a glass of wine, or particular ingredients like spices or alcohol. They can have particular pleasant sensory properties...
like the visual features of a Japanese garden or a gorgeous sunset, the acoustic beauty of Keith Jarrett’s Cologne Concert, the warm feeling of water in the Caribbean, the gorgeous taste of a gourmet dinner, or the irresistible odor of a perfume. Although we need sensory receptors to detect these rewards, their motivating or pleasing properties require further appreciation beyond the processing of sensory components. A good example is Canaletto’s Grand Canal (FIGURE 1) whose particular beauty is based on physical geometric properties, like the off-center Golden Ratio position reflecting a Fibonacci sequence (320). However, there is nothing intrinsically rewarding in this ratio of physical proportions. Its esthetic (and monetary) value is entirely determined by the subjective value assigned by our brain following the sensory processing and identification of asymmetry. Although we process great taste or smell as sensory events, we appreciate them as motivating and pleasing due to our subjective valuation. This rewarding function, cultivated in gourmet eating, enhances the appreciation and discrimination of high-quality and energy-rich primary foods and liquids and thus ultimately leads to better identification of higher quality food and thus higher survival chances (as gourmets are usually not lacking food, this may be an instinctive trait for evolutionary fitness). Sexual attraction is often associated with romantic love that, in contrast to straightforward sex, is not required for reproduction and therefore does not have primary reward functions. However, love induces attachment and facilitates care for offspring and thus supports gene propagation. Sexual rewards constitute also the most straightforward form of social rewards. Other social rewards include friendship, altruism, general social encounters, and societal activities that promote group coherence, cooperation, and competition which are mutually beneficial for group members and thus evolutionarily advantageous.

Nonphysical, nonmaterial rewards, such as novelty, gambling, jokes, suspense, poems, or relaxation, are attractive but less tangible than primary rewards. These rewards have no homeostatic basis and no nutrient value, and often do not promote reproduction directly. We may find the novelty of a country, the content of a joke, or the sequence of words in a poem more rewarding than the straightforward physical aspects of the country or the number of words in the joke or poem. But novelty seeking, and to some extent gambling, may help to encounter new food sources. Jokes, suspense, poems, and relaxation may induce changes of viewpoints and thus help to understand the world, which may help us to consider alternative food sources and mating partners, which is helpful when old sources dry up. Although these rewards act indirectly, they increase evolutionary fitness by enhancing the functions of primary alimentary and reproductive rewards.

Rewards can also be intrinsic to behavior (31, 546, 547). They contrast with extrinsic rewards that provide motivation for behavior and constitute the essence of operant behavior in laboratory tests. Intrinsic rewards are activities that are pleasurable on their own and are undertaken for their own sake, without being the means for getting extrinsic rewards. We may even generate our own rewards through internal decisions. Mice in the wild enter wheels and run on them on repeated occasions without receiving any other reward or benefit, like the proverbial wheel running hamster (358). Movements produce proprioceptive stimulation in muscle spindles and joint receptors, touch stimulation on the body surface, and visual stimulation from seeing the movement, all of which can be perceived as pleasurable and thus have reward functions. Intrinsic rewards are genuine rewards in their own right, as they induce learning, approach, and pleasure, like perfectioning, playing, and enjoying the piano. Although they can serve to condition higher order rewards, they are not conditioned, higher order rewards, as attaining their reward properties does not require pairing with an unconditioned reward. Other examples for intrinsic rewards are exploration, own beauty, gourmet eating, visiting art exhibitions, reading books, taking power and control of people, and investigating the natural order of the world. The pursuit of intrinsic rewards seems private to the individual but may inadver-

![Figure 1](image1.png)

**FIGURE 1.** Subjective esthetic reward value derived from objective physical properties. The beauty of the Canaletto picture depends on the Golden Ratio of horizontal proportions, defined as \((a + b)/a = a/b = 0.618\); \(a\) and \(b\) for width of image. The importance of geometric asymmetry becomes evident when covering the left part of the image until the distant end of the canal becomes the center of the image: this increases image symmetry and visibly reduces beauty. However, there is no intrinsic reason why physical asymmetry would induce subjective value: the beauty appears only in the eye of the beholder. (Canaletto: The Upper Reaches of the Grand Canal in Venice, 1738; National Gallery, London.)
tently lead to primary and extrinsic rewards, like an explorer finding more food sources by venturing farther afield, a beauty queen instinctively promoting attractiveness of better gene carriers, a gourmet improving food quality through heightened culinary awareness, an artist or art collector stimulating the cognitive and emotional capacities of the population, a scholar providing public knowledge from teaching, a politician organizing beneficial cooperation, and a scientist generating medical treatment through research, all of which enhance the chance of survival and reproduction and are thus evolutionary beneficial. The double helix identified by Watson and Crick for purely scientific reasons is now beneficial for developing medications. The added advantage of intrinsic over solely extrinsic rewards is their lack of narrow focus on tangible results, which helps to develop a larger spectrum of skills that can be used for solving wider ranges of problems. Formal mathematical modeling confirms that systems incorporating intrinsic rewards outperform systems relying only on extrinsic rewards (546). Whereas extrinsic rewards such as food and liquids are immediately beneficial, intrinsic rewards are more likely to contribute to fitness only later. The fact that they have survived evolutionary selection suggests that their later benefits outweigh their immediate costs.

D. What Makes Rewards Rewarding?

Why do particular stimuli, objects, events, situations, and activities serve as rewards to produce learning, approach behavior, choices, and positive emotions? There are four separate functions and mechanisms that make rewards rewarding. However, these functions and mechanisms serve the common proximal and distal reward functions of survival and gene propagation. Individuals try to maximize one mechanism only to the extent that the other mechanisms are not compromised, suggesting that the functions and mechanisms are not separate but interdependent.

1. Homeostasis

The first and primary reward function derives from the need of the body to have particular substances for building its structure and maintaining its function. The concentration of these substances and their derivatives is finely regulated and results in homeostatic balance. Deviation from specific set points of this balance requires replenishment of the lost substances, which are contained in foods and liquids. The existence of hunger and thirst sensations demonstrates that individuals associate the absence of necessary substances with foods and liquids. We obviously know implicitly which environmental objects contain the necessary substances. When the blood sodium concentration exceeds its set point, we drink water, but depletion of sodium leads to ingestion of salt (472).

Two brain systems serve to maintain homeostasis. The hypothalamic feeding and drinking centers together with intestinal hormones deal with immediate homeostatic imbalances by rapidly regulating food and liquid intake (24, 46). In contrast, the reward centers mediate reinforcement for learning and provide advance information for economic decisions and thus are able to elicit behaviors for obtaining the necessary substances well before homeostatic imbalances and challenges arise. This preemptive function is evolutionarily beneficial as food and liquid may not always be available when an imbalance arises. Homeostatic imbalances are the likely source of hunger and thirst drives whose reduction is considered a prime factor for eating and drinking in drive reduction theories (242). They engage the hypothalamus for immediate alleviation of the imbalances and the reward systems for preventing them. The distinction in psychology between drive reduction for maintaining homeostasis and reward incentives for learning and pursuit may grossly correspond to the separation of neuronal control centers for homeostasis and reward. The neuroscientific knowledge about distinct hypothalamic and reward systems provides important information for psychological theories about homeostasis and reward.

The need for maintaining homeostatic balance explains the functions of primary rewards and constitutes the evolutionary origin of brain systems that value stimuli, objects, events, situations, and activities as rewards and mediate the learning, approach, and pleasure effects of food and liquid rewards. The function of all nonprimary rewards is built onto the original function related to homeostasis, even when it comes to the highest rewards.

2. Reproduction

In addition to acquiring substances, the other main primary reward function is to ensure gene propagation through sexual reproduction, which requires attraction to mating partners. Sexual activity depends partly on hormones, as shown by the increase of sexual drive with abstinence in human males. Many animals copulate only when their hormones put them in heat. Castration reduces sexual responses, and this deficit is alleviated by testosterone administration in male rats (146). Thus, as with feeding behavior, hormones support the reward functions involved in reproduction.

3. Pleasure

Pleasure is not only one of the three main reward functions but also provides a definition of reward. As homeostasis explains the functions of only a limited number of rewards, the prevailing reason why particular stimuli, objects, events, situations, and activities are rewarding may be pleasure. This applies first of all to sex (who would engage in the ridiculous gymnastics of reproductive activity if it were not for the pleasure) and to the primary homeostatic rewards of food and liquid, and extends to money, taste, beauty, social encounters and nonmaterial, internally set, and intrinsic
Pleasure is a passive reaction that derives from the experience or prediction of reward and may lead to a longer lasting state of happiness. Pleasure as hallmark of reward is sufficient for defining a reward, but it may not be necessary. A reward may generate positive learning and approach behavior simply because it contains substances that are essential for body function. When we are hungry we may eat bad and unpleasant meals. A monkey who receives hundreds of small drops of water every morning in the laboratory is unlikely to feel a rush of pleasure every time it gets the 0.1 ml. Nevertheless, with these precautions in mind, we may define any stimulus, object, event, activity, or situation that has the potential to produce pleasure as a reward.

Pleasure may add to the attraction provided by the nutrient value of rewards and make the object an even stronger reward, which is important for acquiring homeostatically important rewards. Once a homeostatic reward is experienced, pleasure may explain even better the attraction of rewards than homeostasis. Sensory stimuli are another good example. Although we employ arbitrary, motivationally neutral stimuli in the laboratory for conditioning, some stimuli are simply rewarding because they are pleasant to experience. The esthetic shape, color, texture, viscosity, taste, or smell of many rewards are pleasant and provide own reward value independently of the nutrients they contain (although innate and even conditioned mechanisms may also play a role, see below). Examples are changing visual images, movies, and sexual pictures for which monkeys are willing to exert effort and forego liquid reward (53, 124), and the ever increasing prices of paintings (fame and pride may contribute to their reward value). Not surprisingly, the first animal studies eliciting approach behavior by electrical brain stimulation interpreted their findings as discovery of the brain’s pleasure centers (398), which were later partly associated with midbrain dopamine neurons (103, 155) despite the notorious difficulties of identifying emotions in animals.

4. Innate mechanisms

Innate mechanisms may explain the attractions of several types of reward in addition to homeostasis, hormones, and pleasure. A powerful example is parental affection that derives from instinctive attraction. Ensuring the survival of offspring is essential for gene propagation but involves efforts that are neither driven by homeostasis nor pleasure. As cute as babies are, repeatedly being woken up at night is not pleasurable. Generational attraction may work also in the other direction. Babies look more at human faces than at scrambled pictures of similar sensory intensity (610), which might be evolutionary beneficial. It focuses the baby’s attention on particularly important stimuli, initially those coming from parents. Other examples are the sensory aspects of rewards that do not evoke pleasure, are nonnutritional, and are not conditioned but are nevertheless attractive. These may include the shapes, colors, textures, viscosities, tastes, or smells of many rewards (539), although some of them may turn out to be conditioned reinforcers upon closer inspection (640).

5. Punisher avoidance

The cessation of pain is often described as pleasurable. Successful passive or active avoidance of painful events can be rewarding. The termination or avoidance might be viewed as restoring a “homeostasis” of well being, but it is unrelated to proper vegetative homeostasis. Nor is avoidance genuine pleasure, as it is built on an adverse event or situation. The opponent process theory of motivation conceptualizes the reward function of avoidance (532), suggesting that avoidance may be a reward in its own right. Accordingly, the simple timing of a conditioned stimulus relative to an aversive event can turn punishment into reward (584).

E. Rewards Require Brain Function

1. Rewards require brains

Although organisms need sensory receptors to detect rewards, the impact on sensory receptors alone does not explain the effects of rewards on behavior. Nutrients, mating partners, and offspring are not attractive by themselves. Only the brain makes them so. The brain generates subjective preferences that reflect on specific environmental stimuli, objects, events, situations, and activities as rewards. These preferences are elicited by choices and quantifiable from behavioral reactions, typically choices but also reaction times and other measures. Reward function is explained by assuming the notion of value attributed to individual rewards. Value is not a physical property but determined by brain activity that interprets the potential effect of a reward on survival and reproduction. Thus rewards are internal to the brain and based entirely on brain function (547).

2. Explicit neuronal reward signals

Information processing systems work with signals. In brains, the signals that propagate through the circuits are the action potentials generated by each neuron. The output of the system is the observable behavior. In between are neurons and synapses that transmit and alter the signals. Each neuron works with thousands of messenger molecules and membrane channels that determine the action potentials. The number of action potentials, and somewhat their
pattern, varies monotonically with sensory stimulation, discrimination, and movements (3, 385). Thus the key substrates for the brain’s function in reward are specific neuronal signals that occur in a limited number of brain structures, including midbrain dopamine neurons, striatum, amygdala, and orbitofrontal cortex (FIGURE 3). Reward signals are also found in most component structures of the basal ganglia and the cerebral cortical areas, often in association with sensory or motor activity. The signals can be measured as action potentials by neurophysiology and are also reflected in transmitter concentrations assessed by electrochemistry (638) and as synaptic potentials detected by magnetic resonance imaging in blood oxygen level dependent (BOLD) signals (328). Whereas lesions in humans and animals demonstrate necessary involvements of specific brain structures in behavioral processes, they do not inform about the way the brain processes the information underlying these processes. Electric and optogenetic stimulation evokes action potentials and thus helps to dissect the influence of individual brain structures on behavior, but it does not replicate the natural signals that occur simultaneously in several interacting structures. Thus the investigation of neuronal signals is an important method for understanding crucial physiological mechanisms for the survival and reproduction of biological organisms.

3. Reward retina

Neuronal signals in sensory systems originate in specific receptors that define the signal content. However, rewards have no dedicated receptors. Neuronal processing would benefit from an explicit signal that identifies a reward irrespective of sensory properties and irrespective of actions required to obtain it. The signal might be analogous to visual responses of photoreceptors in the retina that constitute the first processing stage for visual perception. To obtain an explicit reward signal, the brain would extract the rewarding component from heterogeneous, polysensory environmental objects and events. A signal detecting the reward properties of an apple should not be concerned with its color unless color informs about reward properties of the fruit. Nor should it code the movement required to obtain the apple, other than assessing the involved effort as economic cost. External visual, somatic, auditory, olfactory, and gustatory stimuli predicting original, unconditioned rewards become conditioned rewards through Pavlovian conditioning. The issue for the brain is then to extract the reward information from the heterogeneous responses to the original and conditioned rewards and generate a common reward signal. Neurons carrying such a signal would constitute the first stage in the brain at which the reward property of environmental objects and events would be coded and conveyed to centers engaged in learning, approach, choice, and pleasure. Such abstract reward neurons would be analogous to the retinal photoreceptors as first visual processing stage (519).

Despite the absence of specific reward receptors, there are chemical, thermal, and mechanical receptors in the brain, gut, and liver that detect important and characteristic reward ingredients and components, such as glucose, fatty acids, aromatic amino acids, osmolality, oxygen, carbon dioxide, temperature, and intestinal volume, filling, and contractions. In addition to these exteroceptors, hormone receptors are stimulated by feeding and sex (24, 46). These receptors are closest to being reward receptors but nevertheless detect only physical, sensory reward aspects, whereas reward value is still determined by internal brain activity.

The absence of dedicated receptors that by themselves signal reward value may not reflect evolutionary immaturity, as rewards are as old as multicellular organisms. Rather valuation separate from physical receptors may be an efficient way of coping with the great variety of objects that can serve as rewards at one moment or another, and of adapting reward value to changing requirements, including deprivation and satiation. Rather than having a complex, omnipotent, polysensory receptor that is sensitive to all possible primary and conditioned rewards and levels of deprivation, however infrequent they may occur, it might be easier to have a neuronal mechanism that extracts the reward information from the existing sensory receptors. The resulting neuronal reward signal would be able to detect rewarding properties in a maximum number of environmental objects, increase the harvest of even rare rewards, and relate their value to current body states, which all together enhance the chance of survival. Such a signal would be an efficient solution to the existential problem of benefitting from the
largest possible variety of rewards with reasonable hardware and energy cost.

III. LEARNING

A. Principles of Reward Learning

1. Advantage of learning

Learning is crucial for evolutionary fitness. It allows biological organisms to obtain a large variety of rewards in a wide range of environments without the burden of maintaining hard-wired mechanisms for every likely and unlikely situation. Organisms that can learn and adapt to their environments can live in more widely varying situations and thus acquire more foods and mating partners. Without learning, behavior for these many situations would need to be pre-programmed which would require larger brains with more energy demands. Thus learning saves brain size and energy and thus enhances evolutionary fitness. These advantages likely prompted the evolutionary selection of the learning function of rewards.

Learning processes lead to the selection of those behaviors that result in reward. A stimulus learned by Pavlovian conditioning elicits existing behavior when the stimulus is followed by reward. Natural behavioral reactions such as salivation or approach that improve reward acquisition become more frequent when a stimulus is followed by a reward. Male fish receiving a Pavlovian conditioned stimulus before a female approaches produce more offspring than unconditioned animals (223), thus demonstrating the evolutionary benefit of conditioning. Operant learning enhances the frequency of existing behavior when this results in reward. Thorndike’s cat ran around randomly until it came upon a lever that opened a door to a food source. Then its behavior focused increasingly on the lever for the food. Thus Pavlovian and operant learning commonly lead to selection of behavior that is beneficial for survival.

Learning is instrumental for selecting the most beneficial behaviors that result in the best nutrients and mating partners in the competition for individual and gene survival. In this sense selection through learning is analogous to the evolutionary selection of the fittest genes. For both, the common principle is selection of the most efficient characteristics. The difference is on the order of time and scale. Selection of behavior through learning is based on outcome over minutes and hours, whereas selection of traits through evolution is based on survival of the individual. Evolutionary selection includes the susceptibility to reward and the related learning mechanisms that result in the most efficient acquisition of nutrients and mating partners.

2. Pavlovian learning

In Pavlov’s experiment, the dog’s salivation following the bell suggests that it anticipates the sausage. A Pavlovian conditioned visual stimulus would have a similar effect. The substances that individuals need to live are packaged in objects or liquids or are contained in animals they eat. We need to recognize a pineapple to get its juice that contains necessary substances like water, sugar, and fibers. Recognition of these packages could be hard wired into brain function, which would require a good number of neurons to detect a reasonable range of rewards. Alternatively, a flexible mechanism could dynamically condition stimuli and events with the necessary rewards and thus involve much less neurons representing the rewards. That mechanism makes individuals learn new packages when the environment changes. It also allows humans to manufacture new packages that were never encountered during evolution. Through Pavlovian conditioning humans learn the labels for hamburgers, baby food, and alcoholic beverages. Thus Pavlovian conditioning touches the essence of reward functions in behavior and allows individuals to detect a wide range of rewards from a large variety of stimuli while preventing run away neuron numbers and brain size. It is the simplest form of learning that increases evolutionary fitness and was thus selected by evolution.

Pavlovian conditioning makes an important conceptual point about reward function. We do not need to act to undergo Pavlovian conditioning. It happens without our own doing. But our behavior reveals that we have learned something, whether we wanted to or not. Pavlov’s bell by itself would not make the dog salivate. But the dog salivates when it hears the bell that reliably precedes the sausage. From now on, it will always salivate to a bell, in particular when the bell occurs in the laboratory in which it has received all the sausages. Salivation is an automatic component of appetitive, vegetative approach behavior.

Although initially defined to elicit vegetative reactions, Pavlovian conditioning applies also to other behavioral reactions. The bell announcing the sausage elicits also approach behavior involving eye, limb, and licking movements. Thus Pavlovian learning concerns not only vegetative reactions but also skeletal movements. Furthermore, Pavlovian conditioning occurs in a large variety of behavioral situations. When operant learning increases actions that lead to reward, the more reward inadvertently conditions the involved stimuli in a Pavlovian manner. Thus Pavlovian and operant processes often go together in learning. When we move to get a reward, we also react to the Pavlovian conditioned, reward-predicting stimuli. In choices between different rewards, Pavlovian conditioned stimuli provide crucial information about the rewarded options. Thus Pavlovian and operant behavior constitute two closely linked forms of learning that extend well beyond laboratory experiments. Finally, Pavlovian conditioning concerns not
only the motivating components of rewards but also their attentional aspects. It confers motivational salience to arbitrary stimuli that elicits stimulus-driven attention and directs top-down attention to the reward and thus focuses behavior on pursuing and acquiring the reward. Taken together, Pavlovian conditioning constitutes a fundamental mechanism that is crucial for a large range of learning processes.

3. Reward prediction and information

Appetitive Pavlovian conditioning takes the past experience of rewards to form predictions and provide information about rewards. The bell in Pavlov’s conditioning experiment has become a sausage predictor for the animal. By licking to the bell the dog demonstrates an expectation of the sausage that was evoked by the bell. Thus Pavlovian conditioning of the bell to the sausage has made the intrinsically neutral bell a reward predictor. The external stimulus or event (the bell) has become a predictor (of the sausage) and evokes an internal expectation (of the sausage) in the animal (FIGURE 4). The same holds for any other reward predictor. The pineapple and the hamburger are predictors of the nutrients they contain.

Predictions tell us what is going to happen. This includes predictors of probabilistic rewards, even if the reward does not occur in every instance. Pavlov’s bell predicts a sausage to the dog and at the same time induces salivation, licking, and approach. Thus Pavlovian conditioning confers two components, a predictive and an incentive property. The predictive component indicates what is going to happen now. The incentive property induces action, such as salivation, licking, and approach, which help to obtain and ingest the sausage. The two properties are separated in time in the classic and widely used delayed response tasks in which initial instructive stimuli have reward-predicting properties without eliciting a reward-directed action (but ocular saccades), whereas the final trigger or releasing stimulus induces the behavioral action and thus has both predictive and incentive properties. The predictive and incentive properties are separated spatially with different stimulus and goal (lever) positions and are dissociable in the behavior of particular rat strains. Specially bred sign-tracking rats approach the conditioned predictive stimulus, whereas goal trackers go directly to the reward upon stimulus appearance, indicating separation of predictive and incentive properties in goal trackers (163).

The predictive component can be further distinguished from an informational component. Once a stimulus has been Pavlovian conditioned, it confers information about the reward. The information does not necessarily predict explicitly what is actually going to happen every time, not even probabilistically. Through Pavlovian conditioning I have learned that a particular sign on a building indicates a pub because I have experienced the beer inside. The atmosphere and the beer represent a value to me that is assigned in a Pavlovian manner to the pub sign. I can pass the pub sign without entering the pub, however difficult that may be. Thus the sign is informational but does not truly predict a pint, only its potential, nor does it have the incentive properties at that moment to make me go in and get one. Then I may run into an unknown pub and experience a different beer, and I undergo another round of Pavlovian conditioning to the value of that particular pub sign. When I need to choose between different pubs, I use the information about their values rather than explicit predictions of getting a beer in every one of them. Thus Pavlovian conditioning sets up predictions that contain reward information. The predictions indicate that a reward is going to occur this time, whereas the reward informations do not necessarily result in reward every time.

The distinction between explicit prediction and information is important for theories of competitive decision-making. Value information about a reward is captured in the term of action value in machine learning, which indicates the value resulting from a particular action, without requiring to actually choosing it and obtaining that value (575). In analogy, object value indicates the value of a reward object irrespective of choosing and obtaining it. Individuals make decisions by comparing the values between the different actions or objects available and then selecting only the action or object with the highest value (see below FIGURE 36C). Thus the decision is based on the information about

![FIGURE 4](http://physrev.physiology.org/) Pavlovian reward prediction. With conditioning, an arbitrary stimulus becomes a reward predictor and elicits an internal expectation of reward. Some of the behavioral reactions typical for reward occur also after the stimulus (Pavlovian stimulus substitution), in particular approach behavior, indicating that the stimulus has acquired reward value (blue arrow).
each reward value and not on the explicit prediction that
every one of these rewards will be actually experienced, as
only the chosen reward will occur. In this way, Pavlovian
conditioning is a crucial building block for reward informa-
tion in economic decisions. Separate from the decision
mechanism, the acquisition and updating of action values
and object values requires actual experience or predictions
derived from models of the world, which is captured by
model free and model-based reinforcement learning, respec-
tively. Other ways to establish reward predictions and in-
formations involve observational learning, instructions,
and deliberate reflections that require more elaborate cog-
nitive processes. All of these forms produce reward infor-
mation that allows informed, competitive choices between
rewards.

4. Pavlovian learning produces higher order rewards

Reward-predicting stimuli established through Pavlovian
conditioning become higher order, conditioned rewards. By
itself a red apple is an object without any intrinsic meaning.
However, after having experienced its nutritious and pleas-
antly tasting contents, the apple with its shape and color has
become a reinforcer in its own right. As a higher order,
conditioned reward, the apple serves all the defining func-
tions of rewards, namely, learning, approach behavior, and
pleasure. The apple serves as a reinforcer for learning to find
the vendor’s market stand. When we see the apple, we ap-
proach it if we have enough appetite. We even approach the
market stand after the apple has done its learning job. And
seeing the delicious apple evokes a pleasant feeling. The
apple serves these functions irrespective of explicitly pre-
dicting the imminent reception of its content or simply in-
forming about its content without being selected. Thus Pav-
lovian conditioning labels arbitrary stimuli and events as
higher order rewards that elicit all reward functions. All this
happens without any physical change in the apple. The only
change is in the eye of the beholder, which we infer from
behavioral reaction.

The notion that Pavlovian conditioning confers higher or-
der order reward properties to arbitrary stimuli and events allows
us to address a basic question. Where in the body is the
original, unconditioned effect of rewarding substances lo-
cated (517, 640)? The functions of alimentary rewards de-
rive from their effects on homeostatic mechanisms involved
in building and maintaining body structure and function. In
the case of an apple, the effect might be an increase in blood
sugar concentration. As this effect would be difficult to
perceive, the attraction of an apple might derive from vari-
ous stimuli conditioned to the sugar increase, such vision of
the apple, taste, or other conditioned stimuli. Mice with
knocked out sweet taste receptors can learn, approach, and
choose sucrose (118), suggesting that the calories and the
resulting blood sugar increase constitute the unconditioned
reward effect instead of or in addition to taste. Besides being
an unconditioned reward (by evoking pleasure or via innate
mechanisms, see above), taste may also be a conditioned
reward, similar to the sensory properties of other alimen-
tary rewards, including temperature (a glass of unpleasant,
luke warm water predicting reduced plasma osmolality) and
viscosity (a boring chocolate drink predicting calories). These
sensations guide ingestion that will ultimately lead to the
primary reward effect. However, the conditioned prop-
erties are only rewarding as long as the original, uncondi-
tioned reward actually occurs. Diets lacking just a single
essential amino acid lose their reward functions within a
few hours or days and food aversion sets in (181, 239, 480).
The essential amino acids are detected by chemosensitive
neurons in olfactory cortex (314), suggesting a location for
the original effects of amino acids with rewarding func-
tions. Thus the better discernible higher order reward fa-
cilitate the function of primary rewards that have much
slower and poorly perceptible vegetative effects.

A similar argument can be made for rewards that do not
address homeostasis but are based on the pleasure they
 evoke. The capacity of some unconditioned rewards to
evoke pleasure and similar positive emotions is entirely de-
termined by brain physiology. Pleasure as an unconditioned
reward can serve to produce higher order, conditioned re-
wards that are also pleasurable. A good example are sexual
stimuli, like body parts, that have unconditioned, innate
reward functions and serve to make intrinsically neutral
stimuli, like signs for particular stores or bars, predictors of
sexual events and activity.

Taken together, the primary, homeostatic, or pleasurable
reward functions are innate and determined by the physi-
ology of the body and its brain that emerged from evolution-
ary selection. Individuals without such brains or with brains
not sensing the necessary rewards have conceivably per-
ished in evolution. The primary rewards come in various
forms and “packages” depending on the environment in
which individuals live. The same glucose molecule is pack-
aged in sugar beets or bananas in different parts of the
world. Conditioning through the actual experience within a
given environment facilitates the detection of the primary
rewards. Thus, in producing higher order rewards, Pavlov-
ian learning allows individuals to acquire and optimally
select homeostatic and pleasurable rewards whose primary
actions may be distant and often difficult to detect.

5. Operant learning

Getting rewards for free sounds like paradise. But after
Adam took Eve’s apple, free rewards became more rare, and
Pavlovian learning lost its exclusiveness. We often have to
do something to get rewards. Learning starts with a reward
that seems to come out of nowhere but actually came from
an action, and we have to try and figure out what made it
appear. The rat presses a lever by chance and receives a drop
of sugar solution. The sugar is great, and it presses again,
and again. It comes back for more. The frequency of its

864 Physiol Rev • VOL 95 • JULY 2015 • www.prv.org
behavior that results in the sugar increases. This is positive reinforcement, the more reward the animal gets the more it acts, the essence of Thorndike’s Law of Effect (589). Crucial for operant learning is that the animal learns about getting the sugar drop only by pressing the lever. Without lever pressing it would not get any sugar, and the behavior would not get reinforced. If the sugar comes also without lever pressing, the sugar does not depend on lever pressing, and the rat would not learn to operate the lever (but it might learn in a Pavlovian manner that the experimental box predicts sugar drops).

Operant reinforcement means that we act more when we get more from the action. However, due to the somewhat voluntary nature of the action involved, the inverse works also. The more we act, the more we get. Our behavior determines how much we get. We get something for an action, which is the bonus or effort function of reward. We can achieve something by acting. We have control over what we get. Knowing this can motivate us. Operant learning allows behavior to become directed at a rewarding goal. We act to obtain a reward. Behavior becomes causal for obtaining reward. Thus operant learning represents a mechanism by which we act on the world to obtain more rewards. Coconuts drop only occasionally from the tree next to us. But we can shake the tree or sample other trees, not just because the coconut reinforced this behavior but also because we have the goal to get more coconuts. In eliciting goal-directed action, operant learning increases our possibilities to obtain the rewards we need and thus enhances our chance for survival and reproduction.

6. Value updating, goal-directed behavior, and habits

Value informations for choices need to be updated when reward conditions change. For example, food consumption increases the specific satiety for the consumed reward and thus decreases its subjective value while it is being consumed. In addition, the general satiety evolving in parallel lowers the reward value of all other foods. Although the values of all rewards ever encountered could be continuously updated, it would take less processing and be more efficient to update reward values only at the time when the rewards are actually encountered and contribute to choices. Reward values that are computed relative to other options should be updated when the value of any of the other option changes.

Operant learning directs our behavior towards known outcomes. In goal-directed behavior, the outcome is represented during the behavior leading to the outcome, and furthermore, the contingency of that outcome on the action is represented (133). In contrast, habits arise through repeated performance of instrumental actions in a stereotyped fashion. Habits are not learned with a new task from its outset but form gradually after an initial declarative, goal-directed learning phase (habit formation rather than habit learning). They are characterized by stimulus-response (S-R) performance that becomes “stamped-in” by reinforcement and leads to inflexible behavior. Despite their automaticity, habits extinguish with repressor omission, as do other forms of behavior. In permitting relatively automatic performance of routine tasks, habits free our attention, and thus our brains, for other tasks. They are efficient ways of dealing with repeated situations and thus are evolutionary beneficial.

Value updating differs between goal-directed behavior, Pavlovian predictions, and habits (133). Specific tests include devaluation of outcomes by satiation that reduces subjective reward value (133). In goal-directed behavior, such satiation reduces the operant response the next time the action is performed. Crucially, the effect occurs without pairing the action with the devalued outcome, which would be conventional extinction. The devaluation has affected the representation of the outcome that is being accessed during the action. In contrast, habits continue at unreduced magnitude until the devalued outcome is experienced after the action, at which point conventional extinction occurs. Thus, with habits, values are only retrieved from memory and updated at the time of behavior. Devaluation without action-outcome pairing is a sensitive test for distinguishing goal-directed behavior from habits and is increasingly used in neuroscience (133, 399, 646). Like goal-directed behavior, Pavlovian values can be sensitive to devaluation via representations and update similarly without repairing, convenient for rapid economic choices (however, Pavlovian conditioning is not goal “directed” as it does not require actions). Correspondingly, reward neurons in the ventral pallidum start responding to salt solutions when they become rewarding following salt depletion, even before actually experiencing the salt in the new depletion state (591). In conclusion, whether updating is immediate in goal-directed behavior or gradually with habit experience, values are only computed and updated at the time of behavior.

B. Neuronal Signals for Reward Learning

1. Basic conditions of learning: contiguity and contingency

In both Pavlovian and operant conditioning, an event (stimulus or action) is paired in time, and often in location, with a reinforcer to produce learning. In delay conditioning, the stimulus or action lasts until the reinforcer occurs. In trace conditioning, which is often less effective (47, 423), the stimulus or action terminates well before the reinforcer. Trace conditioning, but not delay conditioning, with aversive outcomes is disrupted by lesions of the hippocampus (47, 356, 553), although presenting the stimulus briefly again with the outcome restores trace conditioning (27). Thus temporal contiguity is important for conditioning.
What Pavlov did not know was that stimulus-reward pairing was not sufficient for conditioning. Rescorla showed in his “truly random” experiment that the key variable underlying conditioning is the difference in outcome between the presence and absence of a stimulus (or action) (459). Conditioning occurs only when reinforcement is dependent or “contingent” on the stimulus (or an action), and this condition applies also to rewards (126). When the same reward occurs also without the stimulus, the stimulus is still paired with reward but carries no specific reward information, and no conditioning occurs. Thus contingency refers to the intuitive notion that we learn only something that carries specific information or that decreases the uncertainty in the environment (reduction of informational entropy). Contingency is usually defined as difference in reinforcement between stimulus (or action) presence and background (stimulus absence) (FIGURE 5A). A positive difference produces excitatory conditioning (positive learning, left of diagonal), whereas a negative difference produces inhibitory conditioning (negative learning, right of diagonal). Zero contingency produces no conditioning (diagonal line) (but likely perceptual learning). Backward conditioning (reward before stimulus) is ineffective or even produces inhibitory conditioning because the reward occurs without relation to the stimulus. Timing theories of conditioning define contingency as stimulus-to-intertrial ratio of reinforcement rates (175). More frequent reward during the stimulus compared with the no-stimulus intertrial interval makes the reward more contingent on the stimulus. When the same reward occurs later after stimulus onset (temporal discounting), stimulus-to-intertrial reward ratio decreases and thus lowers contingency, although other explanations are also valid (see below). Thus, after contiguity and stimulus-outcome pairing, contingency is the crucial process in conditioning. It determines reward predictions and informations, which are crucial for efficient decisions, and it produces higher order rewards, which drive most of our approach behavior.

**FIGURE 5.** Reward contingency. A: role of contingency in learning. Contingency is shown as reward difference between stimulus presence and absence (background). Abscissa and ordinate indicate conditional reward probabilities. Higher reward probability in the presence of the stimulus compared with its absence (background) induces positive conditioning [positive contingency, triangle]. No learning occurs with equal reward probabilities between stimulus and background (diagonal line, rhombus). Reward contingency applies to Pavlovian conditioning [shown here; reward contingent on stimulus] and operant conditioning [reward contingent on action]. [Graph inspired by Dickinson (132).] B: contingency-dependent response in single monkey amygdala neuron. Top: neuronal response to conditioned stimulus [reward P = 0.9; red] set against low background reward probability (P = 0.0; blue) [triangle in A]. Bottom: lack of response to same stimulus paired with same reward (P = 0.9) when background produces same reward probability (P = 0.9) [rhombus in A] which sets reward contingency to 0 and renders the stimulus uninformative. Thus the neuronal response to the reward-predicting stimulus depends entirely on the background reward and thus reflects reward contingency rather than stimulus-reward pairing. A similar drop in neuronal responses occurs with comparable variation in reward magnitude instead of probability (41). Perievent time histograms of neuronal impulses are shown above raster displays in which each dot denotes the time of a neuronal impulse relative to a reference event [stimulus onset, time = 0, vertical line at left; line to right indicates stimulus offset]. [From Bermudez and Schultz (41).]
2. Neuronal contingency tests

Standard neuronal learning and reversal studies test contiguity by differentially pairing stimuli (or actions) with reward. Although they sometimes claim contingency testing, the proper assessment of contingency requires the distinction against contiguity of stimulus-reward pairing. As suggested by the “truly random” procedure (459), the distinction can be achieved by manipulating reward in the absence of the stimulus (or action).

A group of amygdala neurons respond to stimuli predicting more frequent or larger rewards compared with background without stimuli (FIGURE 5B, top). These are typical responses to reward-predicting stimuli seen for 30 years in all reward structures. However, many of these neurons lose their response in a contingency test in which the same reward occurs also during the background without stimuli, even though stimulus, reward during stimulus and stimulus-reward pairing are unchanged (41) (FIGURE 5B, bottom). Thus the responses do not just reflect stimulus-reward pairing but depend also on positive reward contingency, requiring more reward during stimulus presence than absence (however, amygdala neurons are insensitive to negative contingency). Although the stimulus is still paired with the reward, it loses its specific reward information and prediction; the amygdala responses reflect that lost prediction. Correspondingly, in an operant test, animals with amygdala lesions fail to take rewards without action into account and continue to respond after contingency degradation (399), thus showing an important role of amygdala in coding reward contingency.

Reward contingency affects also dopamine responses to conditioned stimuli. In an experiment designed primarily for blocking (see below FIGURE 6A), a phasic environmental stimulus constitutes the background, and the crucial background test varies the reward with that environmental stimulus. Dopamine neurons are activated by a specific rewarded stimulus in the presence of an unrewarded environmental stimulus (positive contingency) but lose the activation in the presence of a rewarded environmental stimulus (zero contingency) (618). The stimulus has no specific information and prediction when the environmental stimulus already fully predicts the reward. Another experiment shows that these contingency-dependent dopamine responses may be important for learning. Using optogenetic, “rewarding” stimulation of dopamine neurons, behavioral responses (nose pokes) are only learned when the stimulation was contingent on the nose pokes, whereas contingency degradation by unpaired, pseudorandom stimulation during nose pokes and background induces extinction [truly random control with altered p(stimulation|nose poke)] (641).

Taken together, amygdala and dopamine neurons take the reward in the absence of the stimulus into account and thus are sensitive to contingency rather than simple stimulus-reward pairing. They fulfill the crucial requirement for coding effective reward information and prediction and thus may provide crucial inputs to neuronal decision mechanisms. Also, these neuronal signals demonstrate a physical basis for the theoretical concept of contingency and thus strengthen the plausibility of the most basic assumption of animal learning theory.

3. Prediction error learning theory

Contingency, in addition to stimulus-reward pairing (contiguity), is necessary for rewards to induce conditioning. Reward is contingent on a stimulus only when it differs between stimulus presence and absence. When reward is identical, including its time of occurrence, between stimulus presence and absence, it is already predicted by the objects that exist during stimulus presence and absence and therefore not surprising. Hence, the stimulus will not be learned.

The formal treatment of surprise in conditioning employs the concept of prediction error. A reward prediction error PE is the difference between received reward \( \lambda \) and reward prediction \( V \) in trial \( t \)

\[
PE(t) = \lambda(t) - V(t) \tag{1}
\]

This formal definition of “error” extends beyond the colloquial meaning of inaccurate behavior. Prediction errors occur whenever \( Equation 1 \) applies, irrespective of whether they simply occur during behavioral learning or are actually being used for learning.

Animal learning theory aims to explain the role of contingency in conditioning by formalizing how prediction errors update reward predictions from previously experienced rewards (460). The new prediction \( V \) for trial \( t+1 \) derives from the current prediction \( V(t) \) and the prediction error \( PE(t) \) weighted by learning rate \( \alpha \)

\[
V(t + 1) = V(t) + \alpha \cdot PE(t) \tag{2}
\]

Note that \( V \) captures the sum of predictions if several stimuli are present. According to this account, conditioning constitutes error driven learning (FIGURE 6A). When the outcome differs from the prediction, a prediction error proportional to that difference is generated and added to the current prediction after weighting by learning rate \( \alpha \). When \( \alpha \) is set to 1.0, the new prediction is updated in one trial. However, learning rates are usually <1.0, which leads to the typical asymptotic learning curves during which prediction errors decline gradually (FIGURE 6B).

Rewards that are better than predicted generate positive prediction errors and lead to response acquisition, whereas worse rewards generate negative prediction errors and lead to extinction of behavior. When the prediction error becomes zero, no further learning occurs and the prediction errors...
remains stable. Thus learning advances only to the extent to which a reward is unpredicted and slows progressively with increasingly accurate predictions. This formalism views learning intuitively as a change in behavior that occurs when encountering something new or different than predicted, whereas behavior stays the same when everything occurs according to “plan.”

Positive prediction errors direct organisms to maximal rewards, whereas negative prediction errors direct them away from poor outcomes. Obviously, individuals appreciate positive prediction errors and hate negative ones. Optimal behavior for maximizing rewards is characterized by many positive prediction errors and few negative ones.

Contingency is based on prediction errors. Contingency is non-zero, and prediction errors are elicited, when reward differs between stimulus presence and absence. This is the crucial requirement for learning. When reward is identical during stimulus presence and absence, there is no contingency, no prediction error, and no learning.

Specific learning situations benefit from extensions of the basic Rescorla-Wagner rule. Attentional learning rules relate the capacity to learn (associability) to the degree of attention evoked by the stimulus, which depends on degree of surprise (prediction error) experienced in the recent past (prediction error in the current trial cannot contribute to attention to the stimulus, as the error occurs only after the stimulus). Learning rate $\alpha$ is modified according to the weighted unsigned (absolute) prediction error in the recent past ($\lambda$)

$$\alpha = \kappa |\lambda(t) - V(t)|$$

with $\kappa$ as weighting parameter. Thus learning rate increases with large prediction errors, as in the beginning of learning or with large step changes, resulting in strong influences of prediction error and thus fast learning according to Equation 2. In contrast, learning slows with small prediction errors from small changes.

When rewards are drawn from different probability distributions, comparable learning can be achieved by scaling the update component $\alpha \cdot \lambda$ to the standard deviation $\sigma$ of the respective reward distribution by either (438)

$$\alpha^* = \alpha / \sigma \tag{3A}$$

or

$$\lambda(t)^* = \lambda(t) / \sigma \tag{3B}$$

Behavioral tests with measurable prediction errors demonstrate that scaling learning rate $\alpha$ (Equation 3A) by standard deviation may produce constant learning rates with probability distributions differing only in standard deviation (384), which can be formulated as

$$\alpha = \kappa / \sigma \left| \lambda(t) - V(t) \right| \tag{3C}$$

Thus risk may affect learning in two separate ways, by increasing learning rate $\alpha$ via the attentional effect of the prediction error (Equation 3), and by decreasing learning rate or prediction error via division by standard deviation $\sigma$ (Equations 3A and 3B).

The Rescorla-Wagner learning rule models Pavlovian conditioning: changes in outcome lead to prediction errors that lead to changes in prediction and consequently changes in behavior. The learning rule extends to operant conditioning: changes in outcome produce prediction errors that lead to behavioral changes which then result in changes in prediction.

In realistic situations, rewards may be unstable, which results in frequent prediction errors, continuous updating of predictions, and correspondingly varying behavior. Thus learning of a single, stable reward is a special case of, and
formally equivalent to, the general process of updating of reward predictions.

Although the Rescorla-Wagner rule models the acquisition of predictions (V in Equation 2), it does not conceptualize predictions as conditioned, higher order reinforcers. However, higher order reward properties are responsible for most reward functions in learning, approach behavior, and emotional reactions, as outlined above (640). The temporal difference (TD) reinforcement model in machine learning captures the learning functions of higher order rewards (574). Its prediction error includes both unconditioned and higher order reinforcers, without distinguishing between them, as

$$TDPE(t) = [\lambda(t) + \gamma V(t)] - V(t - 1)$$  \hspace{1cm} (4)

The $t$ stands now for time step within a given trial, $\gamma$ is a temporal decay (discounting) factor. A temporal, rather than trial by trial, prediction error is adequate when including higher order reinforcers that derive from reward predictors and necessarily occur at earlier times than unconditioned reinforcers. The term $[\lambda(t) + \gamma V(t)]$ replaces the $\lambda(t)$ term in Equation 1 and incorporates at every time step $t$ the unconditioned reward $\lambda(t)$ and the discounted ($\gamma$ sum $\sum$) of future rewards predicted by $V(t)$ serving as higher order reward. The TDPE includes a higher order prediction error that indicates how much the reward predicted by a conditioned stimulus varies relative to an earlier reward prediction. Thus TDPEs are generated by successive conditioned stimuli to the extent that these stimuli carry different reward predictions than the preceding stimulus.

Substituting PE of Equation 2 by TDPE of Equation 4 leads to the TD learning model

$$V(t + 1) = V(t) + \alpha \times TDPE(t)$$  \hspace{1cm} (5)

Thus TD learning advances backwards from an unconditioned reward to the earliest reward predictor. Updating of the earliest reward prediction employs the prediction value of the subsequent stimulus and does not need to wait for the final, unconditioned reward. Thus each of the conditioned, reward-predicting stimuli serves as a higher order reward, which is biologically plausible by reflecting the frequent occurrence of conditioned rewards and addressing the often difficult distinction between the primary or conditioned nature of natural rewards (640). The use of truly unconditioned rewards in TD learning is straightforward (573). TD reinforcement models are efficient and learn complex tasks involving sequential states, like balancing a pole on a moving platform (32) or playing backgammon (586).

4. Dopamine reward prediction error signal

Most midbrain dopamine neurons show rather stereotyped, phasic activations with latencies of <100 ms and durations of <200 ms following unpredicted food or liquid rewards (FIGURE 7A AND B). The response codes the prediction error, namely, the quantitative difference between received and predicted reward value in each trial $t$, which can be generically be expressed as

$$DaResp(t) \sim \lambda(t) - V(t)$$  \hspace{1cm} (6)

The neuronal response is bidirectional, as a reward that is better than predicted elicits an activation (positive prediction error response) and a reward that is worse than predicted induces a depression (negative error response) in monkeys, rats, and mice (FIGURE 7C). A fully predicted reward draws no response. These responses occur with all-or-none reward prediction errors (321, 365, 410, 503, 517, 521, 524), and with graded, quantitative prediction errors (more or less reward than predicted) (33, 102, 144, 161, 376, 598). Thus the dopamine response codes a reward prediction error (Equation 1) that may constitute a biological instantiation of the reinforcement term in Rescorla-Wagner learning (365), TD learning (374), and spectral timing (71).

The negative dopamine reward prediction error response consists of a depression of activity against low background activity. Adequate assessment requires measuring impulse rate over the maximal length of the depression (34), thus avoiding an almost ungraded negative response (33). Even with a graded negative response (161), the negative error response has less dynamic range than the positive error response. However, the complete cessation of dopamine activity for varying durations may have stronger impact on postsynaptic neurons than graded increases and thus may compensate for the limited dynamic range. Even before it comes to synaptic transmission, voltammetrically measured dopamine concentration changes resulting from negative reward prediction errors are symmetric to the changes resulting from positive errors (204). Thus the dopamine reward prediction error signal seems to be bidirectional.

Dopamine responses show features beyond the Rescorla-Wagner learning rule. About one-third of dopamine neurons show also slower activations preceding reward that vary with the standard deviation (risk) of reward probability distributions (161). The risk response reflects the unsigned (absolute) prediction error and, in keeping with the attentional learning rules (425), may affect learning rate $\alpha$ (Equation 3) and thus serve to adjust learning speed. Furthermore, the prediction error response itself scales with the standard deviation of reward probability distributions (598) according to Equation 3B, which may be driven by the slower activation coding standard deviation and serve to achieve comparable learning with different standard deviations. Third, the prediction error response is sensitive to predictions in model-based learning (see below).

The dopamine response at the time of reward fulfills stringent tests for prediction errors conceptualized by animal...
learning theory (FIGURE 7D). In the blocking test, a test stimulus is blocked from acquiring reward prediction when the reward is fully predicted by another stimulus (zero contingency). If reward does occur after the non-predictive test stimulus, it will produce a positive prediction error, and dopamine neurons are activated by that reward (618). In the conditioned inhibition test, a reward occurring after a stimulus explicitly predicting reward absence elicits a super strong positive reward prediction error (because of a negative prediction being subtracted from the reward, Equation 1), and dopamine neurons accordingly show supranormal activation by the surprising reward (Equation 6) (598). Thus the phasic dopamine responses follow the formal theoretical requirements for prediction error coding.

Reward-predicting stimuli induce phasic activations in most dopamine neurons (60-75%) (FIGURE 7G) (322, 503) and increase correlations between dopamine impulses (256). These responses are gradually acquired through
learning and thus code the prediction V of Rescorla-Wagner and TD models. Correspondingly, voltammetrically measured striatal dopamine release shifts from reward to reward-predicting stimuli during conditioning (117). Acquisition of dopamine responses to conditioned stimuli requires prediction errors, as the blocking test shows. Compounding a novel stimulus with an established reward predictor without changing the reward fails to elicit a prediction error and prevents the stimulus from being learned (FIGURE 8A). A conditioned inhibitor elicits a depressant response instead of an activating response (FIGURE 8B).

Results from tests with sequential reward-predicting stimuli suggest that dopamine responses signal higher order reward prediction errors compatible with TD learning (374). Thus the dopamine responses are well described by extending Equation 6 and using $t$ as time steps, which can be generally expressed as

$$\text{DaResp}(t) \sim [\lambda(t) + \gamma \sum V(t)] - V(t - 1) \quad (7)$$

The dopamine response jumps sequentially backwards from the “unconditioned” liquid reward via stimuli predicting this reward to the earliest reward-predicting stimulus, while losing the response to the predicted intermediate stimuli and the predicted reward (FIGURE 8C) (453, 521). Similar to TD learning, the dopamine responses do not distinguish between primary and higher order reward prediction errors. Only their magnitudes, and fractions of responding neurons, decrease due to temporal discounting, which in TD learning is incorporated as $\gamma$. With several consecutive conditioned stimuli, dopamine neurons code the TD prediction error exactly as the predicted value changes between the stimuli (398). In such tasks, dopamine responses match closely the temporal profiles of TD errors that increase and decline with sequential variations in the discounted sum of future rewards (FIGURE 8D) (144). Thus dopamine neurons code reward value at the time of conditioned and unconditioned stimuli relative to the reward prediction at that moment.

The reward-predicting instruction stimuli in delay tasks have little incentive properties and thus provide a distinction between predictive and incentive stimulus properties (see above). These instructions contain information and elicit saccadic eye movements for acquiring the information contained in the stimulus, but the animal is not allowed to react immediately with a movement towards a rewarded target; it must wait for a later, incentive, movement-triggering stimulus. The consistent dopamine responses to these predictive instructions (FIGURE 8C) (301, 349, 453, 521, 527, 579), and their lack of ocular relationships (527), suggest predominant coding of predictive stimulus properties, without requiring incentive properties. In contrast, in specific sign- and goal-tracking rat strains, striatal voltammetric dopamine responses are stronger to stimuli with incentive compared with reward-predictive properties (163).

These differences may be due to different methods (electrophysiological impulses vs. voltammetric dopamine concentrations) and genetic differences in the specifically bred sign versus goal trackers differentially affecting the dopamine response transfer from reward to predictive stimulus.

Hundreds of human neuroimaging studies demonstrate activations induced by reward related stimuli and reward-related actions in the main reward structures (280, 590). A BOLD neuromagnetic (fMRI) signal reflecting reward prediction error is found in the ventral striatum (351, 390) and constitutes probably the most solid reward response in the brain. It likely reflects the dopamine prediction error signal, as it increases with dopamine agonists and decreases with dopamine antagonists (431). It is detected more easily in striatal and frontal dopamine terminal areas than in mid-brain cell body regions (111, 431, 615), because it presumably reflects summed synaptic potentials (328), which may be stronger after the massive axon collateralization in the striatum. The human reward signal allows also to assess neuronal correlates of positive emotional reward functions, which is intrinsically difficult in animals despite recent efforts (44). With the use of raclopride binding with positron emission tomography (PET), amphetamine-induced dopamine concentrations correlate highly with subjective euphoria ratings in ventral (but not dorsal) striatum (140). BOLD responses in ventral striatum correlate with pleasantness ratings for thermal stimuli (along with orbitofrontal and anterior cingulate activations) (486) and with momentary happiness ratings derived from a combination of certain rewards, gambles, and reward prediction errors (along with anterior cingulate and anterior insula activations) (498). These responses in the dopamine receiving striatum suggest a role of dopamine in signaling pleasure and happiness.

5. Dopamine error signal incorporates model-based predictions

Standard conditioning occurs through prediction errors derived from actually experienced outcomes. However, individuals also learn about contexts, situations, rules, sequences, and conditional probabilities, which may be summarily called models of the world. Knowledge derived from such models can affect the prediction in the neuronal prediction error computation, result in a more appropriate teaching signal, and thus tremendously improve learning.

Model-based learning involves two separate processes, the acquisition and updating of the model, and the influence of the model on reinforcement learning, which likely occurs by influencing predictions of the outcome (139, 575). The acquisition and updating of some of the models likely involves cortical rather than dopamine signals. In contrast, dopamine responses seem to incorporate the predictive information from models once they are established.
In one such experiment, sequences of unrewarded trials lead to higher conditional reward probability with increasing trials (increasing hazard function). The underlying temporal structure constitutes a model in which reward prediction increases progressively. Due to increasing reward prediction, later reward delivery induces increasingly weaker positive prediction errors, and reward omission elicits stronger negative prediction errors. In contrast, model-free reinforcement learning would only consider the past unrewarded trials and hence generate progressively decreasing reward prediction and an opposite pattern of errors. In this experiment, dopamine prediction error responses decrease with increasing numbers of trials since the last reward and thus reflect the increasing reward prediction (FIGURE 8E) (382). In another experiment, dopamine reward prediction error responses adapt to previously learned reward probability distributions (598). The responses scale to expected value and standard deviation and show similar magnitudes with 10-fold differences in reward magnitude. In another example of model-based learning, acquisition of a reversal set allows animals to infer reward value of one stimulus after a single reversed trial with the other stimulus. Accordingly, dopamine prediction error responses reflect the inferred reversed reward prediction (68). In each of these examples, the observed dopamine prediction error responses are not fully explained by the immediate experience with reward but incorporate predictions from the task structure (model of the world). Thus dopamine neurons process prediction errors with both model free and model-based reinforcement learning.
6. Dopamine response in cognitive tasks

Bidirectional dopamine prediction error responses occur also in tasks with elaborate cognitive components. Dopamine neurons, and their potential input neurons in lateral habenula, show bidirectional error responses that signal values derived from differential discounting with conditioned reinforcers (66, 67), although concurrent risk influences on subjective value need to be ruled out. More complex tasks require distinction of many more possible confounds. Early studies used spatial delayed, spatial alternation, and delayed matching-to-sample tasks typical for investigating neurons in prefrontal cortex and striatum and report positive and negative prediction error responses to rewards and reward-predicting stimuli compatible with TD learning models (321, 521, 579). Similar dopamine error responses occur in tasks employing sequential movements (503), random dot motion (389), and somatosensory detection (122). Advanced data analyses from these complex tasks reveal inclusion of model-based temporal predictions into the dopamine error response (382) and compatibility with TD models (144). Variations in stimulus coherence (389) and visual search performance (349) result in graded reward probabilities that induce reward prediction errors that are coded by dopamine neurons, although the error coding may not be easily apparent (426). Thus, in all these high-order cognitive tasks, the phasic dopamine response tracks only, and reliably, the reward prediction error.

7. Dopamine event detection response

The dopamine reward prediction error response is preceded by a brief activation that detects any sufficiently strong event before having identified its reward value (FIGURE 9, A–C, blue areas). The early response component occurs with latencies of <100 ms with all sensory modalities, including loud noises, light flashes, and rapidly moving visual stimuli (229, 322, 527, 563) and evokes dopamine release (117). The activation reflects the sensory impact and physical salience of rewards, punishers, and their predictive stimuli (157, 159, 160). It does not fully reflect motivational salience, as it is absent with negative motivating events of low sensory impact (160, 366), negative reward prediction errors (321, 517, 521, 524, 597, 618), and conditioned inhibitors (597). The detection response covaries inversely with the hazard rate (of temporal event occurrence) and thus codes surprise salience (286) or a temporal prediction error (389) as conceptualized in TD models (574). Thus the early dopamine response component reflects the detection of an event before having identified its value.

This initial response component develops into the graded reward prediction error response, which ranges from depression with strong negative prediction errors to strong activation with positive prediction errors (158, 285, 301, 376, 597). The transition from the first to the second response component depends on the time necessary for assessing the value of the stimulus and thus may vary considerably between behavioral tasks. With easy identification, the transfer takes ~50 ms and fuses the two components, for example, when fixed stimulus positions allow rapid stimulus identification without eye saccades (FIGURE 9, A AND B). Then the two components are fused and may appear as a single response whose identity has long been unclear and may occasionally be labeled as attentional (355, 356). A possible intermediate sensory identification step is either not discernible or is not engaging dopamine neurons. In contrast, the components separate into two separate responses with slower stimulus identification, as seen with random dot motion stimuli requiring >200 ms behavioral discrimination time (FIGURE 9C) (389). Thus the early dopamine activation detects events rapidly before identifying them properly, whereas the subsequent bidirectional response component values the event and codes a proper reward prediction error.

The accurate value coding by the second response component is reflected in the error response at the time of the reward (398, 618). Despite the initial dopamine activation by an unrewarded stimulus, a subsequent reward is registered as a surprise eliciting a positive prediction error re-

---

**FIGURE 8.** Dopamine responses to conditioned stimuli in monkeys. A: stimulus responses of a single dopamine neuron during a blocking test. A pretrained stimulus predicts liquid reward and induces a standard dopamine response (top). During compound training, a test stimulus is shown together with the pretrained stimulus while keeping the reward unchanged (middle left). Thus the reward is fully predicted by the pretrained stimulus, no prediction error occurs, and the test stimulus is blocked from learning a reward prediction. Correspondingly, the test stimulus alone fails to induce a dopamine response (bottom). [From Waelert et al. (618).] B: stimulus responses of a single dopamine neuron during a conditioned inhibition test. The reward normally occurring with the pretrained stimulus (top) fails to occur during compound training with a test stimulus (middle left). This procedure makes the test stimulus a predictor of no reward which correspondingly induces a dopamine depression (bottom). [From Tobler et al. (597).] C: stepwise transfer of dopamine response from reward to first reward-predicting stimulus, corresponding to higher order conditioning conceptualized by temporal difference (TD) model. CS2: instruction, CS1: movement releasing [trigger] stimulus in delayed response task. [From Schultz et al. (521).] D: reward prediction error responses closely parallel prediction errors of formal TD model. Left: sequential movement-reward task. One or two stimuli (white N1, N2) precede the movements leading to reward (orange, 1st, 2nd, 3rd). Right: averaged population responses of 26 dopamine neurons at each sequence step (gray bars, numbers indicate reward probabilities in %) and time course of modeled prediction errors [(f(t) + γV(0) − V(t − 1)) (black line). [From Enomoto et al. (144).] E: dopamine prediction error responses reflect model-based reward prediction derived from temporal task structure. The model specifies an increase in conditional reward probability R(reward | no reward yet) from initial $P = 0.0625$ to $P = 1.0$ after six unrewarded trials. Correspondingly, positive prediction errors with reward occurrence decrease across successive trials, and negative errors with reward omission increase. Averaged population responses of 32 dopamine neurons show similar temporal profiles (blue and red, as impulses/s above neuronal background activity). [From Nakahara et al. (382), with permission from Elsevier.]
sponse (FIGURE 9D, top); accordingly, reward absence fails to induce a negative prediction error response (FIGURE 9D, bottom). Thus, from the second response component on, and before a behavioral reaction can occur, dopamine responses accurately reflect the unrewarded stimulus nature. This constitutes the information postsynaptic neurons are likely to receive, which may explain why animals do not show generalized behavioral responses despite initial neuronal response generalization.

The initial dopamine detection response component varies with four factors. First, it increases with the sensory impact of any physical event, which confers physical salience (FIGURE 10A) (160). This is most likely what the early studies on dopamine salience responses observed (229, 322, 563). Second, also involving memory processing stages, the dopamine detection component is enhanced by stimulus novelty (conferring novelty salience) (322) or surprise (conferring surprise salience) (255, 286, 345). Although novelty may constitute outright reward (458), dopamine neurons are not activated by novelty per se but require stimuli with sufficient sensory impact. The neurons do not respond to small novel stimuli, even though they show prediction error responses to the same small stimuli after learning (597, 618). The total dopamine response decreases with stimulus repetition, together with the animal’s ocular orienting responses, and increases again with reward prediction learning (FIGURE 10B) (517). Third, the dopamine detection component increases with generalization to rewarded stimuli, even when animals discriminate behaviorally well between them. The neuronal generalization is analogous to generalization in behavioral conditioning when an unconditioned stimulus resembles closely a conditioned stimulus and the “associative strength” spills over (335). The incidence of activations to unrewarded, including aversive, stimuli increases with closer similarity and same sensory modality as rewarded stimuli (527, 597, 618). A change from auditory to visual modality of the rewarded stimulus, while leaving the visual aversive stimulus unchanged, increases the initial activations from <15 to 65% of dopamine neurons (FIGURE 10C) (366). Generalization likely also induces motivational salience via similar positively valenced item discrimination of dopamine response into initial detection response and subsequent prediction error response. Increasing motion coherence (from 0 to 50%) improves binary dot motion discrimination and translates into increasing reward probability (from P = 0.49 to P = 0.99). The first response component is nondifferentially constant (blue), whereas the second component grows with increasing reward value (derived from probability, bottom to top, red). [From Nomoto et al. (389).]

Note the initial indiscriminate detection response (CS) predicting reward (gray) and no reward (black). [From Tobler et al. (597).] B: averaged population responses of 54 monkey dopamine neurons to conditioned stimuli (CS) predicting rewards at different delays (2, 4, 8, and 16 s; brown, green, orange, and blue, respectively). The value reduction due to temporal discounting affects only the second, reward prediction error component (red). [From Kobayashi and Schultz (285).] C: differentiation of dopamine response into initial detection response and subsequent prediction error response. Increasing motion coherence (from 0 to 50%) improves binary dot motion discrimination and translates into increasing reward probability (from P = 0.49 to P = 0.99). The first response component is nondifferentially constant (blue), whereas the second component grows with increasing reward value (derived from probability, bottom to top, red). [From Nomoto et al. (389).] D: accurate value coding at time of reward despite initial indiscriminate stimulus detection response. After the unrewarded conditioned stimulus (CS), surprising reward (R) elicits a positive prediction error response (top), whereas predicted reward absence (noR) fails to elicit a negative error response (bottom). [From Waelti et al. (618).]
stimulus responses. During visual search, frontal eye field neurons respond initially indiscriminately to target and distractor and distinguish between them only after 50–80 ms (see below, FIGURE 38B) (588). Neurons in primary visual cortex V1 take 100 ms after the initial visual response to show selective orientation tuning (468) and spatial frequency selectivity (61). Over a time course of 140 ms, V1 neurons initially detect visual motion, then segregate the figure created by dot motion from background motion, and finally distinguish the relevant figure from a distractor (474). V1 and frontal eye field neurons respond initially indiscriminately to targets and distractors and distinguish between them only after 40–130 ms (434). During perceptual decision making, LIP and prefrontal neurons respond initially indiscriminately to stimulus motion and distinguish between motion strength only at 120–200 ms after stimulus onset (see below, FIGURE 38A) (274, 483, 535). Pulvinar neurons show initial, nondifferential detection responses and subsequent differential stimulus ambiguity responses.

FIGURE 10. Four factors influencing the detection component of phasic dopamine responses. A: detection response generated by sensory impact, conferring physical salience. Louder, nonaversive sounds with higher physical salience generate stronger activations (72 and 90 dB, respectively; behavioral choice preferences demonstrated their nonaversive nature). Averaged population responses measured as impulses/s (imp/s) of 14 and 31 monkey dopamine neurons, respectively. [From Fiorillo et al. (160).] B: detection response, and possibly also second response component, enhanced by stimulus novelty, conferring novelty or surprise salience. Stimulus novelty itself is not sufficient to induce dopamine activations, as shown by response absence with small stimuli (horizontal line), but enhances detection response when stimuli are physically larger and more salient (vertical axis). Neuronal responses wane with stimulus repetition due to loss of novelty and increase again with conditioning to reward (from left to right). [Composite scheme from Schultz (517), derived from original data (221, 597, 618).] C: detection response enhanced by generalization to rewarded stimuli. Blue: minor population response to conditioned visual aversive stimulus alternating with auditory reward-predicting stimulus (REW auditory) (active avoidance task). Red: substantial activation to identical visual aversive stimulus when the alternate reward-predicting stimulus is also visual (REW visual), a situation more prone to stimulus generalization. As control, both auditory and visual reward-predicting stimuli induce typical dopamine activations (not shown). [From Mirenowicz and Schultz (366).] D: detection response enhanced by reward context. Left (separate contexts): minor dopamine population activations induced by unrewarded big and small pictures when non-reward context is well separated from reward context by testing in separate trial blocks, using distinct background pictures and removing liquid spout in picture trials. Right (common reward context): major activations by same unrewarded pictures without separation between non-reward and reward context. [From Kobayashi and Schultz (286).]
NEURONAL REWARD AND DECISION SIGNALS

Aversiveness (µL)

180 µL Bitter

10 mM (n=19)

10 mM (n=14)

Bitter

Denatonium

Aversiveness (µL)

Saline (n=8)

Bitter (n=8)

% choosing

0 50 100 150 µL

100

75

50

25

0

0 200 400 600 800 Time after stimulus

Time (ms)

Imp/s

0

2

4

6

8

Bitter Denatonium 180 µl Bitter

FIGURE 11. No aversive coding in monkey dopamine activations. A: psychophysical assessment of aversiveness of bitter solution (denatonium) as prerequisite for investigating quantitative neuronal processing of aversiveness. Monkey chose between single juice reward and juice reward + denatonium (top). The blue and black psychophysics curves show that increasing the volume of single juice reward induces more frequent choices of this reward against constant alternative juice + denatonium. Aversiveness of denatonium is expressed as difference between volume of single juice and juice together with denatonium at choice indifference (50% choice). Thus 1 mM denatonium is worth ~100 µl of juice (black), and 10 mM denatonium is worth ~180 µl (blue). The x-axis shows volume of single juice, and y-axis shows percent of choices of single juice. B: inverse relationship of neuronal activations to psychophysiologically quantified aversiveness of bitter solutions (behavioral method shown in A). N = number of neurons. Imp/s indicate firing rate. C: development of behavioral aversiveness of solutions within individual test days, as assessed with method shown in A. Liquid solutions of salt or bitter molecules lose reward value through gradual satiation and thus become increasingly aversive (thirsty monkeys work for mildly “aversive” saline solutions, and thus find them rewarding, as outcome value is dominated by liquid over salt). D: decreasing dopamine activations with increasing aversiveness of saline solution within individual test days (behavioral aversiveness assessment shown in C), suggesting subtraction of negative aversive value from gradually declining juice value due to satiation. [A–D are from Fiorillo et al. (160).]

Aversive stimuli are well known for >30 years to activate 10–50% of dopamine neurons in awake and anesthetized monkeys, rats, and mice (102, 193, 255, 345, 362, 366, 525, 604) and to enhance dopamine concentration in nucleus accumbens (75). Five neurons close to juxtacellularly labeled ventral tegmental area (VTA) dopamine neurons were activated by footshocks (64), similar to the effects of airpuffs and footshocks on non-dopamine neurons in the area (102, 582). The activations may have derived from dopamine neurons or not; the 30–45% of dopamine neurons lacking spontaneous activity (76, 166, 176, 211) may include nonactivated neurons and thus go undetected while taking up the label. In monkeys, some putative dopamine neurons, often located above substantia nigra, are activated by air puffs (345), although they respond more frequently to conditioned stimuli than primary rewards and thus defy the higher aversiveness of primary than conditioned punishers. However, just as rewards, punishers have distinct sensory, attentional, and motivating components, and none of these studies has distinguished between them. A recent psychophysical study that made this distinction by varying these components independently reports that dopamine activations by aversive stimuli reflect the physical, sensory stimulus impact rather than their aversiveness (FIGURE 11) (160). Once the sensory impact is controlled for, the aversive nature of stimuli either fails entirely to affect dopamine

during visual categorization (288). Tactile responses in mouse barrel cortex require 200 ms to become differential (300). Similar multicomponent responses are seen with reward coding. Amygdala neurons initially detect a visual stimulus and may code its identity and then transition within 60–300 ms to differential reward value coding (9, 422, 428). V1 and inferotemporal cortex responses show initial visual stimulus selectivity and only 50–90 ms later distinguish reward values (371, 557). Thus there is a sequence in the processing of external events that advances from initial detection via identification to valuation. Whereas sensory processing involves only the first two steps, reward processing requires in addition the third step.

B. No aversive dopamine activation

Aversive stimuli are well known for >30 years to activate 10–50% of dopamine neurons in awake and anesthetized monkeys, rats, and mice (102, 193, 255, 345, 362, 366, 525, 604) and to enhance dopamine concentration in nucleus accumbens (75). Five neurons close to juxtacellularly labeled ventral tegmental area (VTA) dopamine neurons were activated by footshocks (64), similar to the effects of airpuffs and footshocks on non-dopamine neurons in the area (102, 582). The activations may have derived from dopamine neurons or not; the 30–45% of dopamine neurons lacking spontaneous activity (76, 166, 176, 211) may include nonactivated neurons and thus go undetected while taking up the label. In monkeys, some putative dopamine neurons, often located above substantia nigra, are activated by air puffs (345), although they respond more frequently to conditioned stimuli than primary rewards and thus defy the higher aversiveness of primary than conditioned punishers. However, just as rewards, punishers have distinct sensory, attentional, and motivating components, and none of these studies has distinguished between them. A recent psychophysical study that made this distinction by varying these components independently reports that dopamine activations by aversive stimuli reflect the physical, sensory stimulus impact rather than their aversiveness (FIGURE 11) (160). Once the sensory impact is controlled for, the aversive nature of stimuli either fails entirely to affect dopamine
neurons (157), reduces the total reward value of liquid solutions and their evoked dopamine activations (157, 160), and induces outright depressant dopamine responses (73, 102, 345, 362, 366, 608). In awake animals, the aversive nature of stimuli does not generate dopamine activations, but induces negative reward prediction errors either because of the negative value or by not being rewarding, and correspondingly induces depressant dopamine responses (157). Thus the different fractions of “aversively” activated dopamine neurons seen in earlier studies likely reflect the different intensities of physical stimuli employed (159). The graded activations with punisher probability (345) might reflect salience differences of conditioned stimuli (286), or graded anticipatory relief to punisher termination (75) that is considered as rewarding (552). Salience, rather than negative reinforcer value, might also explain the surprising higher incidence of activations to conditioned stimuli compared with primary punishment (345). The sensory rather than negative value nature of dopamine responses to aversive stimuli would also explain the dopamine activations by aversive stimuli under anesthesia (64, 525, 604) where motivational impact is ruled out.

The “aversive” dopamine activation constitutes a short latency response that is shorter than the reward response, is substantially boosted by generalization from rewarded stimuli, and is often curtailed by a subsequent depression (366). With these characteristics, the aversive dopamine activations seem to constitute the initial dopamine detection response that reflects mostly physical salience. In agreement with this interpretation, the aversive dopamine neurons are also driven by rewards and unrewarded stimuli and thus are highly sensitive to a large range of stimuli (366) compatible with coding physical impact rather than motivational value. Furthermore, dopamine neurons are neither depressed nor activated by negative aversive prediction errors (157, 255, 345), and show only enhanced activations with positive aversive prediction errors possibly reflecting surprise salience. Thus the current evidence suggests that aversive stimuli activate some dopamine neurons through physical impact rather than through their negative, aversive motivational value, although some dopamine neurons with truly aversive activations can never be completely ruled out.

9. Comprehensive account of phasic dopamine response

Of the two phasic dopamine response components (FIGURE 12), the initial detection activation occurs before the neurons have identified the current reward value of the stimulus and arises even with motivationally neutral events and punishers. It increases with sensory impact, novelty, reward generalization, and rewarded contexts. Thus the initial component detects potential rewards and provides an early report before having identified the value of the event properly, which relates closely to the reward coding of dopamine neurons. The second response component arises once the neurons have well identified the reward value of the stimulus and codes only reward. As longest and strongest component it constitutes the major phasic dopamine influence on postsynaptic processing. The two components are integral parts of the dopamine reward prediction error response, which seems to be split into an early temporal error component and a subsequent value error component (also time sensitive), thus providing an interesting biological instantiation of a TD reinforcement teaching signal (574).

The phasic dopamine reward signal is homogeneous in latency, duration, and sensitivity across different dopamine neurons and very similar between different rewards, thus contrasting sharply with the widely varying reward processing in other brain structures. Graded variations in dopamine responses derive from the initial detection response component that varies in a continuous manner across the mediolateral or dorsoventral extent of the ventral midbrain. If this graded initial component occurs with aversive stimuli and is interpreted as aversive rather than detection response, some dopamine neurons may seem categorically different from those showing a smaller or no detection component (64, 255), suggesting neuronal response heterogeneity (345). However, these variations, as well as variations in the subsequent reward prediction error response (159), conform to a single-peak probability distribution rather than amounting to categorical differences and distinct dopamine subpopulations. In addition to their continuous phasic response distribution, dopamine neurons are het-
The initial activation component may reflect physical, motivational, and surprise salience and thus may enhance the impact of the prediction error response in reward learning compatible with associability learning rules (425, 336). In being highly sensitive to novelty, reward generalization, and rewarded contexts, this component may reflect an early assumption about the nature of an event that is novel, resembles a reward, or occurs within a generally rewarding context. Such events have a chance to be rewards and thus are potential rewards. In responding to such stimuli, dopamine neurons detect a maximal range of potentially rewarding events and objects. Through the very short latency of these responses, dopamine neurons detect these stimuli very rapidly, even before having identified their value, and can rapidly initiate neuronal processes for approach behavior. Once the stimulus has been identified and turns out not to be a reward, it is still time to alter neuronal processing and modify or halt behavioral reactions. If the stimulus is indeed a reward, precious time may be gained and the reward approached before anybody else without such a rapid detection system arrives. By overreacting and processing also potential rewards, the mechanism would prevent premature asymptotes in reward detection and minimize reward misses (avoiding “I do not move unless the reward is for sure”). Such a mechanism would be particularly important when sparse resources present challenges for survival. With early detection facilitating rapid behavioral reactions, the two-component nature of the phasic dopamine response results in better reward acquisition and thus provides competitive evolutionary advantages.

The phasic dopamine reward signal is only one manifestation of dopamine function in the brain. In addition to the distinct and slightly less phasic dopamine risk signal (see below), dopamine exerts a tonic enabling effect on postsynaptic neurons, whose function is considerably more heterogeneous than the rather stereotyped dopamine prediction error response, even when considering its initial detection component (518). Many of these tonic functions depend on the function of the brain structure influenced by tonic extracellular dopamine concentrations, as inferred from behavioral and neuronal alterations arising from inactivations, receptor agonist or antagonist applications, and polymorphisms. These functions extend from motor processes in the striatum evidenced by Parkinsonian movement deficits to prefrontal processes of working memory, attention, reversal learning, categorization, and response control (470). An active involvement of phasic dopamine signals in these vastly different functions is questionable, in particular for functions that are recovered by dopamine receptor agonists without reinstating phasic dopamine signaling. Parkinsonism does not result from deficient dopamine reward prediction error signaling. Thus dopamine function in the brain is vast and heterogeneous and extends well beyond the phasic dopamine reward signal. Serotonin has a similar tonic enabling function, which includes aversive and reward processes (100, 201, 481, 534, 635). The question “what is dopamine doing” reflects the initial idea of “one neurotransmitter-one function” that may originate from “one brain structure-one function.” In view of the many different neuronal dopamine processes, a better question might be “what is the function of the specific dopamine process we are looking at.” For the phasic dopamine signal, the answer seems to be bidirectional reward prediction error coding, with additional, slightly slower risk coding.

10. Dopamine responses during reward learning

During learning episodes, dopamine responses show gradual changes that reflect the evolution of prediction errors. Dopamine neurons acquire responses to reward-predicting stimuli (322) that transfer back to the first reward-predicting stimulus in longer task schedules (144, 521). Their activations disappear gradually with extinction (598), possibly due to inhibition from neighboring pars reticulata neurons activated during extinction (408). When learning several tasks in sequence, positive reward prediction errors are generated every time a reward occurs during initial trials but subside once rewards become predictable and performance approaches asymptote. Dopamine neurons show corresponding prediction-dependent reward responses during these periods (FIGURE 13, A AND B) (322, 521). A temporal difference model using dopamine-like prediction errors replicates well behavioral learning and performance in these tasks (FIGURE 13C) (573). During learning set performance with one rewarded and one unrewarded option, reward is initially predicted with $P = 0.5$. Chance selection of each option leads to positive and negative 0.5 prediction errors, respectively, and to corresponding phasic dopamine activations and depressions during initial learning trials (221). The error responses decrease with increasingly correct choices.

The transfer of dopamine activation from “unconditioned” reward to reward-predicting stimuli constitutes conditioning formalized by the Rescorla-Wagner and TD learning rules. Straightforward implementations of the TD model assume gradual TD error backpropagation via imaginary “internal” stimuli (374, 524). However, biological dopamine neurons do not show such gradual backpropagating responses (162) but transfer in single steps from reward to preceding actual stimuli and show simultaneous responses to both reward and stimuli with intermediate learning steps (410). Dopamine concentration transients in ventral striatum show similar transfer (FIGURE 14A) (117, 598). Biologically plausible TD implementations and spectral timing models achieve the single step response transfer to an earlier stimulus with stimulus and eligibility traces that mark recently active
postsynaptic neurons for plasticity, as originally suggested (FIGURE 14, B–D) (71, 410, 573, 574). The eligibility traces may consist of intracellular calcium changes (154, 637), formation of calmodulin-dependent protein kinase II (236), IP₃ increase in dopamine neurons (203), and sustained neuronal activity in striatum and frontal cortex (522) reflecting previous events and choices in prefrontal neurons (30, 531). Thus use of eligibility traces results in good learning without small step back-propagation via imaginary stimuli (410, 573). These models account also better for the temporal sensitivity of prediction error responses (113, 573).

The acquisition of dopamine responses to conditioned stimuli enhances the strength of excitatory synapses onto dopamine neurons (568), and cocaine administration increases spike-time-dependent LTP in dopamine neurons (19, 318). In elucidating the mechanism, burst stimulation of dopamine neurons, mimicking dopamine reward responses, induces NMDA receptor-dependent LTP of synaptic responses in dopamine neurons (203). Compatible with the temporal characteristics of behavioral conditioning (423, 589), the burst needs to follow the synaptic stimulation by at least 0.5 or 1.0 s, whereas postsynaptic burst omission reverses LTP, and inverse timing induces long-term depres-

FIGURE 13. Dopamine prediction error responses reflect changing reward prediction during learning. A: stepwise learning of spatial delayed response task via intermediate spatial subtasks. Each dot shows mean percentage of correct behavioral performance in 20–30 trials during learning (red) and asymptotic performance (blue). B: positive dopamine reward prediction error responses during learning of each subtask, and their disappearance during acquired performance. Each histogram shows averaged responses from 10–35 monkey dopamine neurons recorded during the behavior shown in A. [A and B from Schultz et al. (521).] C: behavioral learning and performance of temporal difference model using the dopamine prediction error signals shown in B. The slower model learning compared with the animal’s learning of the final delayed response task is due to a single step increase of delay from 1 to 3 s in the model, whereas the delay increased gradually in animals (A). [From Suri and Schultz (573), with permission from Elsevier.]
The 0.5- to 1.0-s delay is required to increase intracellular IP3 as potential stimulus eligibility trace. Thus excitatory synapses onto dopamine neurons seem to mediate the acquisition of dopamine neurons to conditioned stimuli.

11. Non-dopamine prediction error signals

The colloquial meaning of error refers to incorrect motor performance. Climbing fibers of the cerebellum and neurons in superior colliculus and frontal and supplementary eye field...

Bidirectional reward prediction error signals occur in several brain structures. Lateral habenula neurons show bidirectional reward prediction error signals that are sign inverted to dopamine responses and may affect dopamine neurons via inhibitory neurons (FIGURE 15A) (344). Select groups of phasically and tonically firing neurons in the striatum and globus pallidus code positive and negative reward prediction errors bidirectionally (FIGURE 15B) (15, 134, 227, 275, 402). Some neurons in the amygdala display separate, bidirectional error coding for reward and punishment (FIGURE 15C) (36). In the cortex, select neurons in anterior cingulate (261, 531) and supplementary eye field (FIGURE 15D) (551) code reward prediction errors bidirectionally. The bidirectional reward prediction error responses in these subcortical and cortical neurons are straightforwardly appropriate for affecting plasticity at specific postsynaptic neurons for reinforcement learning.

Positive and negative components of reward prediction errors are also coded in select groups of non-dopamine reward neurons. Some neurons in the pedunculopontine nucleus are activated by both predicted and unpredicted rewards and by conditioned stimuli (136, 396, 410) but fail to show depressions to omitted rewards and thus process reward without producing a clear error signal (281). Pedunculopontine neurons with midbrain projections above substantia nigra show bidirectional error responses to reward versus no-reward-predicting cues, but only positive prediction error responses to the reward (228). Positive responses to unpredicted but not predicted rewards occur in noradrenaline neurons as part of their attentional sensitivity (22), and in nucleus basalis Meynert neurons (466). In the striatum, some phasically or tonically firing neurons are influenced by unpredicted reward delivery or unpredicted reward omission (21, 255, 275, 556). Some of these striatal responses are stronger in Pavlovian than operant tasks (17) or occur only after particular behavioral actions (556), suggesting selectivity for the kind of behavior that resulted in the reward. In the amygdala, some neurons are more activated by unpredicted than predicted rewards (36, 40) or show unidirectional, rectified activations with both unpredicted reward delivery and omission (475). In anterior cingulate and dorsolateral prefrontal cortex, activations induced by reward omission or erroneous task performance are known for >30 years (386). In some orbitofrontal neurons, unpredicted rewards outside of tasks elicit activations, whereas reward omission fails to elicit responses (602). Some neurons in anterior and posterior cingulate and supplementary eye field are activated by unpredicted delivery or omission of reward (11, 21, 246, 261, 354, 551) and by unpredicted positive or negative feedback provided by conditioned visual reinforcers (344). Unpredictedness enhances existing reward responses in anterior cingulate cortex (206). The unidirectional activations to unpredicted delivery or omission of reward may reflect surprise salience or rectified positive and negative reward prediction errors. In reflecting surprise salience, they may code the learning rate parameter \( \alpha \) of attentional learning rules (425) according to Equation 3 and serve to adjust the speed of learning (475).

12. Non-dopamine responses during reward learning

Tonically active striatal interneurons (TANs) acquire discriminant responses to reward-predicting auditory stimuli within 15 min of pairing with liquid, and lose these responses within 10 min of extinction (14). Neurons in rat striatum, orbitofrontal cortex, and amygdala acquire discriminant responses to movement-instructing and reward-predicting visual, auditory, and olfactory stimuli during learning and reversal (253, 422, 512). The acquired responses are rapidly lost during extinction (29).

Repeated learning of new stimuli leads to learning set performance in which animals acquire novel stimuli within a few trials (202), thus allowing to record from single neurons.
third group of neurons shows reward expectation activity as learning advances, possibly reflecting exploration. These neurons respond to novel stimuli and acquire reward discriminatory responses during learning. Apparently they code valid reward predictions. To the opposite, other neurons respond initially indiscriminately to all novel stimuli and differentiate as learning advances, possibly reflecting exploration. A third group of neurons shows reward expectation activity that reflects an initial general expectation and becomes selective for rewarded trials as learning progresses (FIGURE 16A). The increases in learning rate with electrical striatal stimulation suggest that some of these striatal neurons may mediate reward learning (639).

In analogy to learning set, repeated reversals of reward predictions of the same stimuli lead to reversal set performance in which animals switch after a single reversed trial.
to the alternative option rather than requiring multiple learning trials, even without having experienced a reward. Prefrontal, orbitofrontal, and amygdala neurons show rapid response loss to previously learned stimuli and acquisition of the currently rewarded stimuli (626), closely corresponding to behavioral changes (FIGURE 16A) (422, 485). The changes occur several trials earlier in striatal neurons compared with prefrontal neurons (421), but are similar between orbitofrontal and amygdala neurons (378). Some striatal and cortical responses reflect the correct or incorrect performance of the previous trial in reversal sets, thus bridging information across consecutive trials (FIGURE 16C) (217).

Both learning sets and reversal sets define task structures and behavioral rules which allow individuals to infer reward values of stimuli without having actually experienced the new values. During reversals, responses in globus pallidus, lateral habenula, and dopamine neurons reflect the new reward value of a target already after the first reversed trial, based solely on inference from the other target (FIGURE 16D) (68). Neuronal responses in dorsolateral prefrontal cortex reflect reward value based on inference from paired associates (FIGURE 16E) (411) or transitivity (409). Neuronal responses in the striatum reflect inference by exclusion of alternative stimuli (409). These responses seem to incorporate the acquired rule into their responses, supposedly by accessing model-based learning mechanisms.

C. Dopamine Implementation of Reward Learning

1. Origin of reward prediction error response

The dopamine reward prediction error signal may derive from two basic mechanisms. The dopamine neurons may receive a complete bidirectional reward prediction error signal from presynaptic input. Alternatively, they may compute the signal from heterogeneous components by subtracting the predicted from the experienced reward at the time of the reward and, in the spirit of TD models, at the time of reward predictors. The computation would require separate inputs for positive and negative errors, both at the time of the final reward and at the time of cues predicting all future rewards.

Dopamine neurons receive main reward information from striatum, amygdala, subthalamic nucleus, pedunculopontine nucleus, rostromedial reticular nucleus, and GABAergic neurons of pars reticulata of substantia nigra (625). They receive homeostatic signals from lateral hypothalamus (464). However, dorsal, medial, and orbital prefrontal cortical areas in monkeys provide probably less direct inputs to dopamine neurons than often assumed (625) but could influence them transsynaptically via the striatum.

Bidirectional reward prediction error response components may arise from lateral habenula signals (FIGURE 15A) (344) which themselves derive from globus pallidus (227) but are sign inverted compared with dopamine responses. The habenula exerts strong and exclusively inhibitory influences on dopamine neurons (95, 250, 344) via GABAergic neurons in the rostromedial reticular nucleus (226, 249). The same habenula neurons show bidirectional punishment prediction error signals opposite to their reward prediction error responses. They respond to positive reward prediction errors in a very similar way as to negative punishment prediction errors (by depression), and to negative reward prediction errors similarly as to positive punishment prediction errors (by activation), thus coding value monotonically across rewards and punishers (346). However, it is unclear why the aversive habenula input would not impact on dopamine neurons (157, 160). Bidirectional reward prediction error inputs may also arise from select neurons in a number of other structures, including the striatum (FIGURE 15B) (15, 134, 275), amygdala (FIGURE 15C) (36), anterior cingulate (261, 531), and supplementary eye field (FIGURE 15D) (551). To make these neurons input candidates for the bidirectional dopamine prediction error signal would require demonstration of their projection to dopamine neurons, which may become technically possible by using selective optogenetic stimulation of error processing neurons that project to dopamine neurons.

Positive dopamine prediction error response components may arise from direct excitatory inputs from pedunculopontine neurons (131) responding to sensory stimuli, reward-predicting cues, and rewards (136, 228, 396, 410). In support of this possibility, inactivation of pedunculopontine neurons reduces dopamine stimulus responses (410). Norepinephrine neurons activated by attentional reward components (22) and nucleus basalis Meynert neurons activated by rewards (466) may also project to dopamine neurons. Different groups of striatal neurons exert activating influences on dopamine neurons via double inhibition (407) and respond to reward-predicting stimuli irrespective of the stimuli being themselves predicted (222) or incorporate reward predictions (17, 21, 255, 275, 402, 556). Positive response components may also derive from amygdala neurons that are activated by rewards irrespective of prediction (422), by unpredicted rewards (40), or by unpredicted reward delivery and omission (475). Reward responses in monkey frontal cortex (11, 21, 206, 246, 261, 347, 354) may reach dopamine neurons via the striatum.

Negative dopamine error responses may arise from GABAergic inputs from VTA, substantia nigra pars reticulata, or striatum (625). Some of these inputs are activated by aversive stimuli (10, 102, 582). Optogenetic activation of pars reticulata GABAergic neurons reduces dopamine impulse activity (613). Inhibition from GABAergic pars-
ticulata neurons showing sustained activations during reward prediction may also result in cancellation of reward activations in dopamine neurons by predicted rewards (102).

The computation of reward prediction error requires a reward prediction to be present at the time of the primary or higher order reward. This prediction may be mediated by the well-known sustained activations preceding rewards and reward-predicting stimuli in neurons of structures projecting mono- or polysynaptically to dopamine neurons, including orbitofrontal cortex (544, 602, 628), dorsal and ventral striatum (18, 205, 215, 523), and amygdala (40) (see below, FIGURE 37, D and E).

2. Postsynaptic influences

The phasic, time specific, reward prediction error reporting dopamine signal is propagated from specific dopamine populations in VTA and substantia nigra pars compacta (SNpc) to nucleus accumbens, dorsal striatum, frontal cortex, amygdala, and other forebrain structures. The widespread and diverging anatomical dopamine projections to postsynaptic structures show limited topography (FIGURE 17A) (177, 251, 342). Dopamine release and postsynaptic influence vary within small regions and are locally finely regulated (92). The tonic local dopamine concentration has separable, powerful enabling effects on a large variety of motor, cognitive, and motivational processes that are chal-
lenged by impaired dopamine function induced by Parkin-
sion’s disease, lesions, and psychopharmacological inter-
ventions (518). Distinction of the tonic function against the 
phasic signal is crucial for adequate conceptualization of 
dopamine’s role in behavior. Importantly, interference of 
dopamine neurotransmission does not allow distinction be-
tween the two functions without special procedures. The 
deficits derive often from impaired tonic dopamine function 
and thus do not match the normal functions conveyed by 
phasic signals.

Dopamine synapses on dendritic spines of postsynaptic 
neurons in striatum and frontal cortex show a triad ar-
rangement with cortical afferents that constitutes an archi-
tecture for three-factor Hebbian plasticity and allows the 
global dopamine reinforcement signal to exert differential 
influences on selectively active corticostriatal neurotrans-
mission (FIGURE 17B) (171, 186, 461, 517). Indeed, elec-
trical midbrain stimulation induces dopamine D1 receptor 
dependent long-term potentiation in striatal neurons in 
vivo (461) and expansion of cortical auditory fields for 
costimulated frequencies (28). Dopamine plays crucial 
roles in long-term potentiation (LTP) and depression 
(LTD) in striatum (82, 265, 294, 583), frontal cortex 
(194, 400), hippocampus (401), and amygdala (495).
Dopamine turns tetanus stimulation-induced cortical 
LTD into LTP when paired with NMDA receptor activa-
tion (343). In spike-time-dependent plasticity (STDP) 
protocols, LTP occurs differentially when presynaptic 
stimulation precedes postsynaptic stimulation by a few 
tens of milliseconds, whereas LTD occurs with reverse 
sequence (48). Slice STDP protocols induce anatomical 
enlargement of striatal dendritic spines by burst-stimu-
ulating dopamine axons 0.3–2.0 s after depolarization of 
medium spiny striatal neurons, involving dopamine D1, 
but not D2, receptors and striatal NMDA receptors 
(642). The specific sequence of striatal excitation fol-
lowed, but not preceded, by dopamine stimulation cor-
responds well to the temporal requirements for behav-
ioral conditioning (423, 583), frontal cortex (194, 400), 
hippocampus (401), and amygdala (495).
Dopamine turns tetanus stimulation-induced cortical 
LTD into LTP when paired with NMDA receptor activa-
tion (343). In spike-time-dependent plasticity (STDP) 
protocols, LTP occurs differentially when presynaptic 
stimulation precedes postsynaptic stimulation by a few 
tens of milliseconds, whereas LTD occurs with reverse 
sequence (48). Slice STDP protocols induce anatomical 
enlargement of striatal dendritic spines by burst-stimu-
ulating dopamine axons 0.3–2.0 s after depolarization of 
medium spiny striatal neurons, involving dopamine D1, 
but not D2, receptors and striatal NMDA receptors 
(642). The specific sequence of striatal excitation fol-
lowed, but not preceded, by dopamine stimulation cor-
responds well to the temporal requirements for behav-
ioral conditioning (423, 583), frontal cortex (194, 400), 
hippocampus (401), and amygdala (495).

FIGURE 18. Differential learning impairment by dopamine recep-
tor blockade. A: dopamine receptor blockade induces severe learn-
ing deficits in sign-tracking rats (learning under flupenthixol, not 
shown; test drug free, shown here). Sign trackers contact a condi-
tioned stimulus before approaching the goal. B: dopamine receptor 
blockade by flupenthixol induces less severe learning deficits in goal-
tracking rats. Goal trackers bypass the stimulus and directly ap-
proach the goal. [A and B from Flagel et al. (183). Reprinted with 
permission from Nature Publishing Group.] C: knockout (KO) of 
NMDA receptors on mouse midbrain dopamine neurons reduces 
phasic impulse bursts (not shown) and induces learning deficits in 
pseudorandomly alternating T-maze reward arms. [From Zweifel et 
al. (650).]
play more important roles when prediction errors have more direct effects.

Genetic NMDA receptor knockout on mice dopamine neurons reduces neuronal burst responses to reward-predicting stimuli and, in parallel, induce deficits in a wide range of learning situations including conditioned place preference, maze learning, and operant conditioning (FIGURE 18C) (624, 649, 650). Correspondingly, GABA<sub>A</sub> receptor knockout on dopamine neurons enhances dopamine release and T-maze and lever press learning (414). Without permitting adaptation during ontogenetic development, inactivation of dopamine neurons by local muscimol impairs learning from unexpected reward increase or omission (478). Transient direct or transsynaptic optogenetic inhibition of dopamine neurons induces place dispreference learning in choices between two compartments (see FIGURE 19E below) (244, 582).

Taken together, learning depends on intact dopamine function in simple reward contiguity situations with explicit, easily identifiable rewards and conditioned stimuli that engage phasic dopamine responses. Learning may not depend on dopamine neurons in situations in which they do not show phasic responses. Learning may also occur despite dopamine impairments when other learning systems compensate within the employed time frames, which may consist of switching to other struc-

**FIGURE 19.** Effects of dopamine stimulation on behavioral learning. A: operant self-stimulation chamber. Pressing the lever elicits an electrical or optogenetic stimulus delivered to a specific brain structure through an implanted electrode/optrode. [From Rentato M. E. Sabatini, The history of electrical stimulation of the brain, available at http://www.cerebromente.org.br/n18/history/stimulation_i.htm.] B: place preference conditioning by phasic but not tonic optogenetic activation of dopamine neurons in mice. [From Tsai et al. (605). Reprinted with permission from AAAS.] C: operant nosepoke learning induced by optogenetic activation of dopamine neurons at stimulated target (blue) but not at inactive target (black) in rats. [From Witten et al. (641), with permission from Elsevier.] D: optogenetic activation of dopamine neurons unblocks learning of visual stimulus for nosepoke in rats, as shown by stronger response to stimulus paired with optogenetic stimulation in a blocking procedure (blue) compared with unpaired stimulus (orange) and stimulation in wild-type animals (gray). Response decrments are typical for unreinforced tests (“in extinction”). [From Steinberg et al. (562). Reprinted with permission from Nature Publishing Group.] E: place dispreference conditioning by direct optogenetic inhibition of dopamine neurons in mice (yellow; optical stimulation alone, gray). [From Tan et al. (582), with permission from Elsevier.] F: operant conditioning of approach (blue) and avoidance (red) behavior by optogenetic activation of D1 and D2 dopamine receptor expressing striatal neurons, respectively, in mice. [From Kravitz et al. (293). Reprinted by permission of Nature Publishing Group.]
tures after receptor blockade or lesions or altering ontogenetic development in gene knockouts. Fine-grained behavioral analysis may nevertheless reveal subtle remaining learning deficits. Given the crucial importance of reward for survival, multiple, flexible systems that sustain basic forms of learning despite partial impairments are biologically plausible and would enhance evolutionary fitness.

4. Sufficient dopamine involvement in learning (stimulation)

Intracranial electrical self-stimulation (398) induces operant learning of lever pressing (FIGURE 19A). The effect is partly based on activation of VTA and SNpc dopamine neurons (103, 155) and involves dopamine release onto D1 receptors in nucleus accumbens (117). Optogenetic activation of midbrain dopamine neurons in rodents elicits learning of place preference, nose poking and lever pressing, and restores learning in a blocking procedure (FIGURE 19, B–D) (2, 244, 271, 562, 605, 641). Optogenetically induced nose poke learning occurs also with activation of dopamine axons in rat nucleus accumbens and is attenuated by intra-accumbens infusion of dopamine D1 and D2 receptor antagonists, suggesting involvement of dopamine projections to nucleus accumbens (562). The stimulations are similar in frequency to natural dopamine activations (20–50 Hz) but often exceed their natural durations of 150–250 ms by 2–7 times (500–1,000 ms), although stimulation durations of 200 ms are also effective (271, 641). The longer-than-natural activations induce more dopamine release in vitro (641) but may be compromised by the limited capacity of dopamine neurons for prolonged discharges due to pronounced tendency for depolarization block (54, 463). In contrast to 50-Hz phasic optogenetic dopamine activation, tonic (1 Hz) stimulation does not induce place preference learning (605), emphasizing the efficacy of phasic activations. In monkeys, electrical 200 Hz/200 ms VTA microstimulation induces choice preferences and neuromagnetic striatal activations, presumably reflecting dopamine activations (20). Opposite to excitations of dopamine neurons, direct optogenetic inhibition of dopamine neurons, or their indirect inhibition via activation of local, presynaptic GABA neurons, leads to place dispreference learning (FIGURE 19E) (244, 582). Conceivably the optogenetic excitation and inhibition of dopamine neurons mimic positive and negative dopamine prediction error signals and affect learning accordingly.

Learning induced by stimulation of dopamine neurons and their postsynaptic targets demonstrates the acquisition of associations with stimulation reward rather than simply reflecting arousal, response bias, direct effects on behavior, or performance enhancement, as the learned behavior is retained over several days, declines only gradually during extinction (271, 293, 582, 605) and is sensitive to devaluation by contingency degradation (641). The learning may involve striatal dopamine receptors, as optogenetic stimulation of D1 receptor expressing striatal neurons induces approach learning and stimulation of D2 receptor containing striatal neurons induces avoidance learning (FIGURE 19F) (293). The dichotomous striatal functions go back to learning models showing choice preference mediated by D1, and dispreference by D2, receptor containing striatal neurons and their respectively associated direct and indirect pathways (169). Taken together, electrical and optogenetic activation of dopamine neurons seem to mimic the natural dopamine prediction error responses to rewards and elicit learning behavior. The wider question is, to which extent would dopamine responses to a natural reward bias Pavlovian associations with that reward or operantly condition actions leading to the reward?

IV. APPROACH AND CHOICE

Rewards induce approach behavior and serve as arguments for economic choices. These functions involve a fundamental requirement, the prediction of reward that elicits reward expectation. We can approach an object because we expect it to be a reward. Without that expectation, we would approach an unknown object only when we are exploring the world. The same holds for choices. Final outcomes are often not apparent at the time of choices. Without such information, choices would be guesses and thus be very inefficient. One simply cannot make informed decisions without knowing what to expect from each option. Explorations are not informed choices but are helpful for detecting potentially better rewards. Thustgoal-directed approach behavior and informed choices are based on predictions of reward.

Predictions are acquired in the most basic form by Pavlovian conditioning. The sound of the bell predicts the sausage to Pavlov’s dog. Also, the stimuli occurring during operant conditioning become Pavlovian conditioned reward predictors as the improving behavior results in more rewards. Each stimulus may become a higher order reward through conditioning and learning and thus has distinct economic value. Money does not have homeostatic or reproductive functions on its own but allows individuals to acquire nutritional and mating rewards and thus is a non-primary reward. More complex forms of prediction learning involve observation of other individuals, inference of events based on acquired knowledge, as in Bayesian updating in which the “posterior” probability of event occurrence arises from the “prior” probability, and conditional and reflective reasoning (“if I do this I can expect to obtain that reward”). In mediating Pavlovian and operant conditioning, in addition to stimulus-reward and action-reward pairing, the three-term contingency under which a discriminative stimulus signals an effective action-reward relationship constitutes the key factor for goal-directed approach behavior and informed choices. Contingency requires prediction errors, which serve to acquire and update the crucial predictions. Reward neurons are sensitive to contingency (FIGURE 5) (41) and code reward prediction errors (FIGURE 7) (517).
A. Basic Reward Processing

1. Events eliciting reward responses

Explicit neuronal reward signals code only reward information. They occur as phasic responses to the delivery of “unconditioned” rewards, as phasic responses to conditioned, reward-predicting stimuli, and as sustained activity during the expectation of reward (FIGURE 20A). They reflect only reward properties and do not vary with sensory or motor aspects (FIGURE 20, B AND C).

“Unconditioned” liquid or food rewards elicit explicit reward signals in all main components of the brain’s reward system (see FIGURES 7; 8, A–C, 13B, AND 27A) (16, 18, 42, 60, 205, 215, 322, 383, 387, 405, 422, 451, 492, 541, 592). Reward responses are found also in premotor, prefrontal, cingulate, insular, and perirhinal cortex (8, 246, 369, 569). Some prefrontal neurons code the reward of the preceding trial (30, 179).

Conditioned, reward-predicting stimuli induce phasic explicit reward signals in all main reward structures of the brain, including orbitofrontal cortex, striatum, amygdala, and dopamine neurons (VTA and SNpc) (FIGURE 20, B AND C; see also FIGURES 5B; 7C; 8C; 9, A, B, AND D; 27, B AND D) (e.g., Refs. 41, 205, 276, 319, 371, 387, 405, 422, 516, 542, 601). These responses occur irrespective of the physical properties of the conditioned stimuli. Similar reward-related stimulus responses occur also in dorsolateral prefrontal cortex and anterior insula (338, 369, 411).
During reward expectations evoked by reward-predicting stimuli, select neurons in orbitofrontal cortex, striatum, and amygdala show slower, sustained explicit reward signals (see Figures 16A AND 38, D AND E) (18, 37, 215, 601) which may also reflect reward timing (528) and temporal reward structure (40).

In addition to inducing explicit reward signals, rewards affect existing sensory-specific and action-specific activity. Neurons coding reward together with differential sensory information are found in orbitofrontal, dorsolateral prefrontal, perirhinal, and inferotemporal cortex (231, 319, 371, 394, 411). Reward affects neuronal activity differentiating between different movement parameters during the instruction, preparation and execution of action in prefrontal and premotor cortex (Figure 20, D AND E) (135, 282, 313, 348, 476, 606, 627), anterior and posterior cingulate cortex (354, 542), parietal cortex (381, 427, 433), striatum (Figure 20F) (107, 205, 222, 260, 308), substantia nigra pars reticulata (502), superior colliculus (243), and amygdala (428). By processing information about the forthcoming reward during the preparation or execution of action, these activities may reflect a representation of the reward before and during the movement toward the reward, which fulfills a crucial requirement for goal-directed behavior (133). However, in motor structures, increased movement-related activity with larger rewards may reflect the more energized movements rather than representing a true conjoint reward-action signal (476). After reward delivery, responses in dorsolateral prefrontal neurons differentiate between movement directions (606).

Taken together, reward neurons show passive responses to reward-predicting stimuli and rewards, and sustained activities in advance of predicted rewards. Explicit reward signals in a limited number of brain structures reflect reward information but code neither sensory nor motor information. In contrast, sensory and motor neurons in several brain structures process reward information conjointly with sensory responses and action activity (Figure 21). Reward coding involves the large majority of dopamine neurons, which all respond in a very similar manner to the effective events. In contrast, only specific and much smaller fractions of neurons in the other reward structures show reward activity, and their responses can be quite selective for one or the other of several task events.

2. Subjective reward perception

Neuronal reward responses may depend on the subjective perception of reward-predicting stimuli. In a signal detection task, monkeys correctly or incorrectly report the presence of stimuli (hit and false alarm, respectively), and correctly or incorrectly report stimulus absence (correct rejection and miss) (122). Dopamine neurons are activated by an initial task stimulus only when the animal correctly reports its detection, whereas they are not activated by the same physical stimulus when the animal reports its absence (miss). Thus subjective perception of the stimulus is necessary to activate dopamine neurons. However, subjective perception is not sufficient, as dopamine neurons do not respond when the physically absent stimulus is incorrectly reported as present (false alarm) (Figure 22). The prediction error responses to a subsequent cue and to the final reward reflect the subjective perception of the initial
stimulus. The responses are lower, and thus incorporate the prediction from the initial stimulus, when the animal had accurately reported stimulus presence compared with inaccurately reporting stimulus absence, even though the stimulus was physically the same. Thus dopamine responses reflect the animal’s subjective perception and valuation of reward-predicting stimuli beyond their purely physical properties.

3. Distinction to unrewarding events

Reward neurons differentiate reward against neutral and aversive events. Dopamine neurons code reward but are not activated by aversive stimulus components (FIGURE 11) (157, 159, 160), even though their initial event detection response covaries with physical salience, novelty salience, stimulus generalization, and reward context (FIGURE 10). In contrast, neurons in other reward structures respond selectively to aversive stimuli. Non-dopamine VTA neurons are activated by airpuffs and footshocks (102, 582). Some amygdala neurons are activated by airpuffs and conditioned stimuli predicting airpuffs, separately from reward neurons (422). Neurons in nucleus accumbens are differentially activated by sucrose or quinine, including their predictive stimuli (484). Neurons in orbitofrontal and anterior cingulate cortex, and TANs in striatum, respond differentially to aversive airpuffs, reward liquids, and their predictive stimuli and follow reversal between punishment and reward (10, 255, 452, 592). In dorsolateral prefrontal cortex, more neurons respond to reward than airpuff (283) and are differentially activated by gains or losses of reward tokens (532). In medial prefrontal cortex, dorsal neurons respond more frequently to air puffs, whereas ventral neurons are more sensitive to liquid reward (372). Thus punishment responses are often well separated from reward processing.

Some reward neurons respond also to punishers and thus process the motivational salience common to both reinforcers rather than reward or punishment separately. These reinforcer neurons are seen in amygdala (387), dorsolateral prefrontal cortex (283), and parietal cortex (309). In contrast to separate monotonic variations with reward and punishment magnitude, salience responses peak with high reward and high punishment and thus follow a U function (309). Thus restricted populations of reward neurons fail to discriminate against punishers and thus code motivational salience.

However, there is a caveat when interpreting the results of studies using aversive stimuli for probing negative motivational value. As indicated for dopamine responses above, it is crucial to distinguish between the physical impact of the stimulus and its motivational aversiveness. However, most studies apply presumably aversive stimuli without assessing their aversiveness to the individual animal. A good method to do so in neurophysiological studies is to estimate the negative value quantitatively on a common currency basis, using psychophysical choice procedures against rewards (FIGURE 11A) (160). Even substantial airpuffs are surprisingly not reported by the animals as having negative reward value (160).

B. Reward Value

1. Central role of subjective reward value

The ultimate function of reward is evolutionary fitness, which requires us to survive and reproduce. Thus rewards have survival value that is specific for us, and thus subjective. “Subjective” does not refer to conscious awareness but simply to the own person. Our brain detects the rewards that are necessary for this function. It is not really important what the objective, physical amount of a reward is, and often not even possible to know, as long as the reward provides us with the best chance for survival and reproduction. If we have plenty of food, a bit more does not matter much, but if we are starving, even a small amount is important. We do not necessarily go for the largest reward if it does not mean anything for us. We go for the reward that most satisfies our needs at that moment. A rat normally readily does not drink a salt solution but will do so when it is deprived of salt (472). Striving for the highest subjective value is built into our behavior by evolutionary selection. If we do not choose the highest value, we are at a disadvantage and over time fall behind and ultimately lose out against competitors that get better values. The striving for subjective value might also explain why we sometimes make choices that others do not understand. They are good for us at the moment, because we make them based on our own, private, subjective values, irrespective of whether we are consciously aware of them or not. Subjective valuation is also mechanistically reasonable. Reward function in individuals is defined by behavior driven by the brain, and the brain has evolved for our individual survival and reproduction rather than for an impersonal cause. We can identify all
substances in a wine and add up their prices from a catalog, but as a chemist and not as a wine connoisseur who is trying to maximize the personal pleasure based on subjective taste preferences. It is not the chemists’ impersonal cause directed at the physical molecules but the wine connoisseur’s subjective pleasure that contributes to fitness (by providing relaxation and encouraging taste discrimination training).

2. Assessing subjective value from behavior

Although subjective value is the key variable underlying approach and choices, it is a theoretical construct and not an objective measure. Utilitarianism aimed to achieve the greatest happiness for the greatest number of people (39). This might be assessed by an external person with honorable intentions (the benevolent dictator), by a hypothetical hedonimeter that measures happiness (142), or by subjective ratings (for example as “experienced utility”) (259). However, benevolent dictators are implausible, hedonimeters are technically difficult, and ratings are subjective. Furthermore, there is no way of comparing subjective values between individuals because we cannot compare their feelings or, as a famous economist put it, “You cannot feel how much my tooth aches.” In contrast, behavioral reactions and choices are measurable. They are often described in terms of preferences. However, preferences are internal private states and therefore prone to misjudgement and misreport. To operationalize preferences and remove their metaphysical and emotional connotations of purpose, desire, and pleasure, Samuelson (501) coined the term revealed preferences that are elicited by choices and follow specific axioms, notably consistency and transitivity. Thus choices elicit private and unobservable preferences hidden deep inside us and make them measurable as utility in an objective way. Other, more loosely defined, terms for value include survival value (survival of individuals or genes, referring to the reward function in evolutionary fitness), motivational or incentive value (eliciting approach behavior), and affective value (linking reward to emotion). There will be a more specific definition of economic utility further down with neuronal value signals.

There are several mechanisms for inferring subjective value from approach behavior and choices. In approach behavior, reward value is revealed by the speed (e.g., reaction time), duration, intensity, frequency, accuracy (e.g., error rate), and persistence of behavioral reactions. The reward eliciting the strongest approach behavior presumably has the highest “value.” In choices, reward value is revealed by the probability of choosing one option over all other alternatives. The reward being consistently selected is considered to have the highest subjective value (assuming “rational” choices). In animals, psychophysical assessment of choices provides good measures for value. In humans, auction mechanisms allow to assess the subjective value of goods (willingness to pay). Thus approach and choices reveal the unobservable variable of value. When I see some-one move faster towards an object or choose the object over another one, I get the intuitive notion that the object must have a higher value for the person. Value becomes simply a shortcut term denoting the impact of a reward on measurable behavioral choices. To avoid the circular definition when using value for explaining behavior we say that individuals behave “as if” they are maximizing value. Deducing the variable of value from behavior allows us to search for neuronal value signals for approach and choices. Such neuronal value signals would provide a physical basis for the theoretical notion of value and thus inform and validate economic choice theory.

3. Relative choices and common currency value

Behavioral choices reflect the values of choice options relative to each other. They will remain stable irrespective of value variations as long as the relative values of the options remain unchanged. For example, doubling of all option values should not affect choices. Thus value can only be expressed relative to other options within a given set of options. We can define a common reference reward, compare behavioral choices between that reference reward and any other reward, and thus establish a common value ranking or numeric scale defined by the reference reward which then serves as common currency. Money is a good example. It assesses the value of any reward in common monetary units. Any reward can be used as common currency for rewards differing in type, magnitude, probability, risk, delay, and other attributes. Just as different physical objects are measured on common scales of length or weight, different rewards can be measured on a common scale of value. Note that the common scale is only valid for an individual decision maker and does not provide a common measure to all decision makers. One dollar, or any other common currency, may constitute a different value for you than for me.

How can we choose between apples and oranges? We can compare the sensory properties of different objects using standardized scales of physical and chemical measures. As reward value is not a physical but a behavioral measure, this does not help. Rather, we should assess the values of such incommensurable rewards relative to each other. The idea of common currency allows us to quantify value relative to a unique reference that quantifies the value of rewards with vastly different properties. The activity of orbitofrontal and dopamine neurons reflects the subjective, common currency value integrated from different liquid rewards in close correspondence to behavioral choices (405) and extends to food rewards (301) (Figure 23), suggesting a physical basis for the theoretical notion of common currency. Common currency is a necessary and sufficient feature of value coding, as it reflects independence from sensory properties and follows the idea of a reward retina as earliest neuronal detector of value (519).
Relative values do not always correlate with choices. Decision makers have a tendency to select the highest option unproportionally more often than its relative value against other options. Such maximizing choices favor the highest valued option irrespective of value variations as long as the ordinal value ranking among the options remains unchanged. Maximizing is reduced during exploratory behavior, when individuals occasionally check an inferior option for potential improvement, but exploration declines with increasingly stable values. These mechanisms are captured by the “temperature” parameter of the softmax function that models choices (333, 575). Exploration, modeled by higher temperature, may lead to effective decorrelation between choice probability and value and reveal serious constraints on the computation of value from choices. However, specific test methods control for these possible misreadings, including assessment of indifference points (subjective equivalence between choice options) in psychological procedures.

4. Value in natural rewards

Natural rewards, for which reward processing has evolved, contain nutrients and other substances of survival value. These rewards, like foods and drinks, have sensory properties, like shape, color, texture, viscosity, taste, and smell that help to detect, identify, and distinguish the reward objects. However, the sensory properties do not have nutrient value. Approach behavior and choices maximize reward value, not sensory properties. The transfer from sensory properties to value involves experiencing the value of the object, associating that value with the object’s sensory properties using Pavlovian and higher forms of learning, and then using the predicted value for approach and choices. Thus sensory discrimination serves to identify the object via sensory receptors rather than determining its value. The sensory-value distinction lies at the core of reward processing; the sensory properties belong to the external object, whereas the reward properties derive from internal brain function attributing value to external objects.

Reward neurons distinguish between the sensory and value attributes of rewards. Amygdala and dopamine neurons show faster sensory and slower value responses (160, 422). Orbitofrontal and amygdala neurons follow reversed value associations of physically unchanged stimuli (422, 485). Orbitofrontal and striatal reward neurons discriminate between different food and liquid rewards but are insensitive to their visual and spatial attributes (205, 405, 601). Satiation on particular rewards reduces reward value responses in orbitofrontal and ventromedial prefrontal cortex (58, 105) but leaves sensory responses in primary gustatory cortex unaffected (487). Salt depletion induces reward responses in ventral pallidum to physically unchanged salt solutions (591). Thus the intuitive distinction between sensory and value attributes of rewards has a neurophysiological basis.

5. Non-value factors in choices

Besides being based on positive and negative economic value (gain and loss and their corresponding emotions) and exploration, real life economic choices are also determined by other factors. The factors include strategy, heuristics,
personal history, peer example, social pressure, conventions, traditions, prejudice, idiosyncrasies, morals, ethics, culture, religion, superstition, nationalism, chauvinism, daily irrationalities, and many others. Some originate from experiences that maximized gains and minimized losses and then led to automatisms that are no longer questioned for economic value. (Note that such automatisms are distinct from habits which may develop from goal-directed behavior and are based on value.) Other choice factors have evolved without explicit understanding of distant gains. Just as birds do not know the partial differential equations describing the coordination of their movements, individuals engage in behaviors without understanding why they do them, how they do them, and what the benefits are. These automatisms have evolved through evolution and provide guidance of behavior to varying but often large extents. Their advantages are transmission of beneficial behavioral traits, but their possible disadvantages are inconsistencies and domination by higher values during choices. Decisions based on automatisms would involve brain processes related to schemata and models of behavior, rules, emotions, social memory, theory of mind, and other cognitive functions.

The impact of automatisms should be considered in relation to value as key variable for economic choices. The true character of most automatisms differs fundamentally from that of gains and losses, and decision makers may not compute economic values when following habitual strategies or conventions. Then deriving value from economic choices would be very difficult and needs to be restricted to specific situations in which the automatism component can be estimated. This is often difficult and may easily lead to inconsistent choices when different situations engage different automatisms. What is truly missing is a definition of value independent of choice behavior. A neuronal value signal might provide such a measure and allow interpretations of economic choices. For example, a neuronal value signal identified independently of a particular choice situation might help to assess the contribution of value in choices in which automatisms also play a role.

C. Construction of Neuronal Value Signals

1. Value functions

Intuitively, the value of a reward increases monotonically with its physical magnitude. A larger drop of juice is worth more than a smaller drop, except when satiation results in asymptotic and then decreasing value and thus breaks monotonicity (see below). The value of a reward depends also on the frequency of its occurrence which is modeled as probability. Stimuli that predict more likely rewards are valued higher. Given their occurrence with specific magnitudes and probabilities, rewards constitute probability distributions of individual reward magnitudes, and choices between two rewards can be viewed as choices between probability distributions (417). The first statistical moment in probability distributions is the expected value (EV), which constitutes the sum of probability-weighted magnitudes (first moment of probability distributions)

$$\text{EV} = \sum_i [m_i \cdot p_i]; \text{ over all } i \quad (8)$$

with $m$ as magnitude (objective property of good), $p$ as objective probability, and $i$ as individual reward occurrences. The mean of increasingly large samples approaches the EV.

Probability is a theoretical construct that is derived from the measured frequency of event occurrence. Probability theory, as developed by Pascal and Fermat, allowed transferring the measure of experienced event frequency into the predictive measure of probability. For biological agents and their neurons, this transfer requires learning and memory, as the frequency of events impinging on sensory receptors needs to be counted in relation to other events and expressed as probability. The fact that there are no sensory receptors for probability strengthens the notion of a theoretical construct. Thus probability assessments mediate the transition from reacting to individual events to predicting future event occurrence. The past experience of reward frequency is registered by neurons and, true to the nature of predictions, processed to predict rewards and guide future behavior. By presenting events with a given frequency to animals, we assume that this transfer takes place, and we can search for neuronal activity that reflects reward probabilities. As described above with Pavlovian conditioning, the activity is necessarily predictive of the overall frequency of reward occurrence and thus reflects the reward value of the predictive stimulus. Thus the brain reacts to and assesses the frequency of individual rewards and transfers this measure into predictive activity reflecting the probability of future rewards.

The intuition of subjective value underlying economic choices combines with the failure to explain human choices by EV (43) and led to the axiomatic formulation of Expected Utility Theory (504, 617), which defines expected utility as a measure of subjective value in analogy to Equation 8 as

$$\text{EU} = \sum_i [u(m_i) \cdot p_i]; \text{ over all } i \quad (9)$$

with $u(m)$ as (subjective) utility. The main utility functions are power, logarithmic, exponential, and quadratic as well as their combinations (295), which can model human choices adequately (225). Logarithmic utility may combine the logarithmic Weber function that describes the sensory impact of reward objects with a logarithmic valuation function of objective, physical reward magnitude and probability.
Utility is a hypothetical variable reflecting the subjective preferences elicited by behavioral choices. Whereas objective, physical value is measurable directly in milliliters or grams, utility is not measurable directly but inferred from behavioral choices (and maximized by decision makers). Thus the only physical measure from which utility can be inferred is behavioral choices. Choices are the tool to assess utility, and choices are the key method that economists accept for this assessment. There is a crucial difference between subjective value and utility. Although both subjective value and utility are estimated from measured behavioral choices, utility is a mathematical function of objective value \( u(m) \) that allows to predict the subjective value even for choices that have not been measured. Such mathematical functions allow to determine whole distributions of subjective values and establish useful terms such as EU for comparison between choice options. Utility is a universal measure of subjective value that does not require immediate behavioral assessment every time and thus constitutes the fundamental variable of economic decision theory.

The curvature of utility functions reflects marginal utility. Marginal utility is defined as the increment in utility gained from one additional unit of consumption, or the decrement from one unit less consumption, which is mathematically the first derivative of the utility function. Progressively decreasing marginal utility leads to concave functions (downward concave, as viewed from below, decreasing first derivative). The probability weighting function \( \pi(p) \) is most nonlinear at 0.0 < \( p \) < 0.2 and \( 0.5 < p < 1.0 \) and can be modeled with a one- or two-parameter function (\( \alpha \)) (435) with an inflection point \( \pi(p) = p \) at about \( P = 0.37 \), or a linear-in-log-odds function (187) in which \( \gamma \) reflects curvature (low = inverted S, high = regular S) and \( \alpha \) reflects elevation. Low probabilities can be distorted by a factor of 10,000. Well-trained monkeys show similar nonlinear, inverted S-shape probability weighting as humans with similar inflections points around \( P = 0.37 \) (FIGURE 25) (559).

Given that rewards affect behavior primarily through their subjective value, and that utility \( u(m) \) constitutes the central subjective reward function in economics, we can use Equations 9 and 10 to state reinforcement learning. Equations 1, 2, 4, and 5 in terms of economic utility

\[
\text{UPE}(t) = \lambda(t) - \text{EU}(t) \quad (10A)
\]

\[
\text{EU}(t + 1) = \text{EU}(t) + \alpha \ast \text{UPE}(t) \quad (10B)
\]

\[
\text{UTDPE}(t) = \left[ (\lambda(t) + \gamma \ast \text{EU}(t)) \right] - \text{EU}(t - 1) \quad (10C)
\]

\[
\text{EU}(t + 1) = \text{EU}(t) + \alpha \ast \text{UTDPE}(t) \quad (10D)
\]

The subjective weighting of value extends to reward probability. In particular, low and high probabilities are often distorted. Prospect Theory refines Equation 9 to

\[
\text{EU} = \sum_i [u(m_i) \ast \pi(p_i)] \quad \text{over all } i \quad (10)
\]

with \( \pi \) as probability weighting function (and \( P \) replacing EU in popular notation) (258) (Prospect Theory in addition incorporates references, different curvatures, and different slopes for losses than gains). The probability weighting function \( \pi(p) \) is most nonlinear at 0.0 < \( p \) < 0.2 and \( 0.5 < p < 1.0 \) and can be modeled with a one- or two-parameter function (\( \alpha \)) (435) with an inflection point \( \pi(p) = p \) at about \( P = 0.37 \), or a linear-in-log-odds function (187) in which \( \gamma \) reflects curvature (low = inverted S, high = regular S) and \( \alpha \) reflects elevation. Low probabilities can be distorted by a factor of 10,000. Well-trained monkeys show similar nonlinear, inverted S-shape probability weighting as humans with similar inflections points around \( P = 0.37 \) (FIGURE 25) (559).

Given that rewards affect behavior primarily through their subjective value, and that utility \( u(m) \) constitutes the central subjective reward function in economics, we can use Equations 9 and 10 to state reinforcement learning. Equations 1, 2, 4, and 5 in terms of economic utility

\[
\text{UPE}(t) = \lambda(t) - \text{EU}(t) \quad (10A)
\]

\[
\text{EU}(t + 1) = \text{EU}(t) + \alpha \ast \text{UPE}(t) \quad (10B)
\]

\[
\text{UTDPE}(t) = \left[ (\lambda(t) + \gamma \ast \text{EU}(t)) \right] - \text{EU}(t - 1) \quad (10C)
\]

\[
\text{EU}(t + 1) = \text{EU}(t) + \alpha \ast \text{UTDPE}(t) \quad (10D)
\]

The subjective weighting of value extends to reward probability. In particular, low and high probabilities are often distorted. Prospect Theory refines Equation 9 to

\[
\text{EU} = \sum_i [u(m_i) \ast \pi(p_i)] \quad \text{over all } i \quad (10)
\]

with \( \pi \) as probability weighting function (and \( P \) replacing EU in popular notation) (258) (Prospect Theory in addition incorporates references, different curvatures, and different slopes for losses than gains). The probability weighting function \( \pi(p) \) is most nonlinear at 0.0 < \( p \) < 0.2 and \( 0.5 < p < 1.0 \) and can be modeled with a one- or two-parameter function (\( \alpha \)) (435) with an inflection point \( \pi(p) = p \) at about \( P = 0.37 \), or a linear-in-log-odds function (187) in which \( \gamma \) reflects curvature (low = inverted S, high = regular S) and \( \alpha \) reflects elevation. Low probabilities can be distorted by a factor of 10,000. Well-trained monkeys show similar nonlinear, inverted S-shape probability weighting as humans with similar inflections points around \( P = 0.37 \) (FIGURE 25) (559).

The “bliss point” is the utility producing maximal satisfaction. In contrast, increasing marginal utility leads to convex functions (FIGURE 24A). A “well-behaved” economic utility function is concave, continuous, monotonically increasing, and nonsaturating (295, 341). It is based on money of which apparently one cannot get enough. Utility functions vary between different rewards, different risk levels (631), and different individuals. Utility functions for other rewards, like foods and drinks, do saturate and may even decrease, as too much food or liquid may become aversive and thus have negative marginal utility, effectively destroying monotonicity of the utility function (FIGURE 24B). The “bliss point” is the utility producing maximal satisfaction. In contrast, increasing marginal utility leads to convex functions (FIGURE 24C), suggesting that wins in high ranges are considered more important than the same gains in lower ranges (“small change”). In some cases, gains are valued increasingly more but ultimately become less important, leading to an initially convex and then concave utility function (173, 339) (FIGURE 24D). Effectively, all combinations of curvatures of utility functions are possible, but the typical, “well-behaved” concave utility function is the standard model in economics.

### FIGURE 24.

Utility functions. **A:** a “well behaved,” gradually flattening (“concave”), continuous, monotonically increasing, nonsaturating utility function, typically for money. Gray rectangles indicate decreasing marginal utility \( y \)-axis with increasing wealth despite identical changes in objective value \( x \)-axis. **B:** a saturating, nonmonotonic utility function typical for many biological rewards. Marginal utility becomes negative when an additional unit of consumption reduces utility (right end of function). **C:** a convex utility function, with gradually increasing marginal utility. **D:** a convex-concave utility function typical for progressively increasing marginal utility in lower ranges and progressively decreasing marginal utility in higher ranges.
Note that $\lambda$ refers to the utility $u(m)$ of the actual single reward received, whereas EU derives from the statistical utility distribution. Thus the utility prediction errors $UPE(t)$ and $UTDPE(t)$ constitute the key reinforcement terms for decision variables defined by economic utility.

Economists distinguish three principal forms of utility. First, scalar utility is subjective value as a mathematical function of physical, objective value. Utility can be derived from marginal utility under risk (to result in cardinal utility) or from the marginal rate of substitution without risk (amounting to ordinal utility). Cardinal utility functions established by choices under risk (617) provide a quantitative, numeric, mathematical characterization of the preference structure (257). Cardinal utility is valid up to positive affine transformations ($y = ax + b$) and can be psychophysically estimated from certainty equivalents at choice indifference between certain and risky options, starting at the distribution ends and advancing centrally by iteration (fractal or chaining procedure) (87). Cardinal utility can be expressed as $u(m)$ (Equation 9) and is a requirement for establishing neuronal utility functions as neuronal signals have cardinal characteristics. In contrast, ordinal utility provides only a rank order of utilities and indicates whether a good is better or worse than another, without indicating on a numeric scale how much better or worse it is. It is only of limited value for establishing neuronal correlates of economic utility (higher versus lower neuronal responses for better or worse goods without numeric scaling). The marginal rate of substitution indicates the amount of good A a consumer is willing to give up to obtain one unit of good B while maintaining the same utility. The “distance” between curves of constant utility (indifference curves) indicates at least ordinal utility. Second, mean-variance utility is derived from objective value and risk via Taylor series expansion (see below, Equations 21 and 21A). Third, prospect theory extends utility by incorporating probability distortion (Figure 25 and Equation 10), different loss than gain slopes (see below Figure 29E) and reference dependency (see below Figure 32 and Equation 22A).

The term of utility derives from the original utilitarian philosophers (39, 363). In economic decision theory, utility is defined by formal axioms (504, 617) that postulate how individuals behave in a meaningful manner “as if” they were maximizing utility. The term of utility is often used for convenience without requiring formal axiomatic tests in every instance (150, 324). This article follows this tradition, although that usage should not discourage axiomatic testing when identifying neuronal utility signals. Thus economic utility is a specific form of subjective value, which is derived from objective behavioral measures in individuals (but not across individuals). Utility has stricter requirements than subjective value by assuming an underlying, usually nonlinear, mathematical function that derives utility from objective value [$u(m)$, as used in Equation 9]. If assessed from risky choices, utility is unique up to a positive affine transformation, which defines a cardinal function.

2. Neuronal value coding

The definition of cardinal, quantitative, numeric utility is important for investigating neuronal utility signals. The prime neuronal signal that is communicated to other neurons is the action potential. With sensory information, the signal arises from dedicated receptors whose stimulation induces numeric neuronal signals. Its strength as a signal is quantitatively expressed as firing rate (impulses/s) (3) and, to some extent, as pattern of action potentials. Firing rate is quasi-continuous between 0 and 1,000 impulses/s within the usual periods of 10–60 min of data sampling and usually increases or decreases monotonically with the phenomenon it codes. The same neuronal processing principles could be used for coding reward utility. For such neuronal signals to approach a meaningful mathematical function of utility, utility needs to be cardinal. It would be incorrect to derive a cardinal neuronal function from ordinal utility. This requirement would not apply if one would only look for ordinal higher firing rate for a higher (or lower) ranked, ordinal reward. With these characteristics, neuronal utility signals are physical (hardware) manifestations of a hypothetical variable that can only be inferred from behavioral choices. Although choices are the key method for assessing utility from behavior, utility neurons should also reflect utility at the time of reward reception, as this is the most important event for the survival of the organism.

Reward magnitude exerts a physical impact on sensory receptors. Reward magnitudes induce monotonically graded responses in dopamine neurons and in neurons in the stria-
temporal discounting of reward value (see below), differently delayed conditioned reinforcers (66, 67), or deprivation without physical reward changes (58, 105, 284, 332, 403, 601). Details will be presented in the following sections.

Stimuli predicting specific reward probabilities induce graded responses in dopamine, striatum, and posterior parietal neurons and also in orbital, ventromedial, ventrolateral, dorsolateral, and cingulate prefrontal neurons, along with graded behavioral reactions and choices (12, 15, 17, 161, 261, 262, 354, 372, 376, 381, 402, 418, 433, 500). The transfer of dopamine responses from rewards to reward-predicting stimuli (FIGURES 7C AND 14, A AND D) may constitute a suitable correlate for the transition from reporting experienced reward frequency to coding reward probability. There are no neurophysiological studies on probability distortions. Human neuroimaging has shown probability distortions of values analogous to Equation 10 (241, 595) but no behavioral distortions with valueless probabilities (110). Thus reward neurons transform the theoretical construct of probability into another theoretical construct called value (both of which are measurable from behavioral choices).

Increasing EVs induce graded neuronal responses in dopamine, striatum, orbitofrontal, ventromedial prefrontal, ventrolateral prefrontal, dorsolateral prefrontal, cingulate, and posterior parietal cortex neurons (FIGURE 26) (12, 261, 262, 354, 381, 402, 433, 418, 500, 598). In addition to characterizing reward coding, these data suggest a general biological basis for the mathematical constructs of EV, for which there are no sensory receptors. However, it is unlikely that these neurons code objective EV specifically rather than the subjective value that is monotonically related to objective value. Nevertheless, neuroscience applauds to the genius of the earlier mathematicians postulating biologically implemented theoretical terms.

Appropriate behavioral tests reveal that reward neurons typically code subjective value. Reward responses in orbitofrontal cortex, ventromedial prefrontal cortex, parietal cortex, striatum, globus pallidus, amygdala, lateral habenula and dopamine neurons closely reflect subjective behavioral choices (FIGURE 23) (301, 405, 433). Subjective value coding is also evident in specific situations, including satiation or deprivation without physical reward changes (58, 105, 591), differently delayed conditioned reinforcers (66, 67), temporal discounting of reward value (see below FIGURE 28D) (79, 158, 285, 477, 479), and adaptation to identical reward magnitudes (see below FIGURE 33, A AND B) (42, 284, 332, 403, 601). Details will be presented in the following sections.

Although subjective value coding confirms the close behavioral-neuronal relationships of reward processing, it is agnostic about the relationship to physical reward value. In contrast, economic utility defines this relationship mathematically. Neuronal correlates of utility are found in dopamine neurons (560). Utility functions in monkeys may transition from convex via linear to concave with increasing reward volumes (FIGURES 24D AND 27A), which reflects nonmonotonic marginal utility (first derivative of utility) and results in nonmonotonic utility prediction errors. This utility profile is closely paralleled by nonmonotonically changing dopamine prediction error responses with gambles placed at different parts of the utility function (FIGURE 27B). With unpredicted rewards eliciting positive prediction errors, dopamine responses increase monotonically (FIGURE 27C) and thus follow the monotonic increase in utility prediction error (Δu) but not the nonmonotonic variation in marginal utility (Δu/Δx or du/dx). Thus the dopamine reward prediction error response constitutes a utility prediction error signal that reflects marginal utility but does not explicitly code marginal utility. Marginal utility relates well to prediction error, as both terms conceptualize deviations from a reference (prediction and current wealth, respectively).

Given its coding of utility, the dopamine signal can be stated as
DaResp(t) \sim\frac{\text{H11011}}{\text{UPE}(t)}(10E)

and formulated for temporal difference (TD) utility error

\text{DaResp}(t) \sim \text{UTDPE}(t) \quad (10F)

and can be fully expressed by using Equations 6 and 7, replacing V by EU and applying Equations 10A and 10E

\text{DaResp}(t) \sim \lambda(t) - \text{EU}(t) \quad (10G)

\text{DaResp}(t) \sim [\lambda(t) + \gamma\sum\text{EU}(t)] - \text{EU}(t-1) \quad (10H)

As some neurons in other reward structures code also reward prediction errors, DaResp in Equations 10E through 10H may be replaced by error signals from such neurons if their utility coding were established in future experiments.

As utility is not simply a measure of subjective value but derives mathematically from objective reward measures, utility coding extends the dopamine response well beyond subjective reward value coding seen with multiple rewards, effort cost, and risk (116, 156, 301, 570). Utility as higher specification of the dopamine response likely applies also to multiple rewards, effort cost, and risk whose neuronal coding is so far characterized only as subjective value (risk is already incorporated into the utility response in Figure 27B).

3. Motivational state

Based on the intuition that the primary function of nutrient rewards derives from the need to acquire necessary substances for survival, reward value is closely related to the state of the animal. Deprivation on a particular substance enhances its subjective value, whereas satiation reduces its value. These states can be general or specific for a substance (e.g., “general” versus “sensory specific” satiety).

Satiation is captured by decreasing marginal utility in gradually flattening (concave) utility functions (FIGURE 24A). Higher satiation leads to complete saturation and subsequent repulsion characterized by zero and then negative marginal utility (FIGURE 24B), although such satiety does not typically occur with money. In contrast, reward deprivation is captured by the steeper part of the utility function with higher marginal utility.
Rather than directly affecting marginal utility and EU, an alternative account may conceptualize motivational states as likely nonlinear and nonmonotonic utility (EU\textsubscript{state}) which adds to EU and results in overall EU\textsubscript{net}, similar to other influences described below

\[
\text{EU}_{\text{net}} = \text{EU} + \text{EU}_{\text{state}} \quad (11)
\]

With deprivation, EU\textsubscript{state} is positive and increases EU\textsubscript{net}, whereas satiation turns EU\textsubscript{state} negative and decreases EU\textsubscript{net}. We can distinguish general from sensory specific effects on EU\textsubscript{net} by using separate state utilities

\[
\text{EU}_{\text{net}} = \text{EU} + \alpha \times \text{EU}_{\text{genstate}} + \beta \times \text{EU}_{\text{specstate}} \quad (11A)
\]

with weighting coefficients $\alpha$ and $\beta$. General deprivation and satiety affect all rewards to some extent (EU\textsubscript{genstate}), whereas sensory specific satiety affects primarily the particular reward (EU\textsubscript{specstate}). The effects of deprivation and satiety on EU\textsubscript{net} can be accommodated by the reinforcement framework by substituting EU in Equations 10A through 10D by EU\textsubscript{net}.

Satiety on individual rewards reduces behavioral reactions and neuronal responses in monkey orbitofrontal cortex to the smell, taste, and sight of rewards (105) and in ventromedial prefrontal cortex to water (58). The response decrease is specific for the satiating reward, including blackcurrant, cream, vanilline, limonène, or monosodium glutamate liquids (105). Some responses to the nonsatiating rewards remain uninfluenced, suggesting sensory specific rather than general satiety. General satiety over the daily course of the experiment reduced most responses of dopamine neurons to auditory cues for food pellets in mice (496). Opposite to satiation, salt depletion makes salt rewarding. Rats show behavioral approach to NaCl solutions after, but not before, salt depletion by furosemide while reward neurons in the ventral pallidum begin to respond to salt (472, 591). These results provide neuronal correlates for the influence of motivational states on subjective value compatible with Equation 11.

4. Context

Rewards occur always in an environmental context, and these contexts exert considerable influence on subjective reward value. Imagine yourself in a gourmet restaurant and hearing a loud “wow.” Then imagine watching a tennis match and hearing the same “wow.” Obviously the behavioral reaction to the “wow” differs depending on the context, suggesting that information from the context (restaurant versus tennis) affects the interpretation of the explicit event (“wow”). Adding a nonlinear subjective value function EU\textsubscript{cxt} captures the value gained by rewarded contexts

\[
\text{EU}_{\text{net}} = \text{EU} + \text{EU}_{\text{cxt}} \quad (12)
\]

Animal learning theory conceptualizes the influences of context on behavioral reactions in several forms. In general, Pavlovian-instrumental transfer, instrumental (operant) learning is enhanced in the presence of a previously Pavlovian conditioned stimulus compared with an unconditioned stimulus (145). In renewal, previously extinguished conditioned responding may return depending on the contextual environment (59). In pseudoconditioning, a primary reinforcer endows the context and all its cues with reward value through conditioning. Pavlov noted “a conditioned reflex to the environment” (423) and Konorski mentioned “conditioned to situational cues” (289). Pseudoconditioning elicits behavior to unpaired events in rewarded contexts (367, 536, 537), as if reward properties “spill over” from the context to the explicit event. For example, in an experiment on rabbit jaw movement, the value of the unpaired stimulus was close to 0 (analogous to EU = 0), but amounted to 70% of the value of the rewarded stimulus after that stimulus had been conditioned in the same context (EU\textsubscript{cxt} = 70% of rewarded stimulus EU) (537). In all these forms, otherwise ineffective stimuli occurring in the same context gain motivational value and induce behavioral responses (191, 335), even though they have never been explicitly paired with the primary reinforcer.

Influences of contextual information are well known from everyday experiences. Labeling an odor as cheese rather than body odor, or describing flavors as rich and delicious rather than vegetable water, increases pleasantness ratings and responses in human orbitofrontal cortex (118, 189). Stating the brand name of soft drinks, like Coke and Pepsi, strongly increases preference ratings for the drink indicated and enhances prefrontal reward responses (352). In wine tasting, stated prices influence pleasantness ratings and value responses in orbitofrontal cortex (432). In risky decisions with only partly known outcomes, additional information of outcomes after the decision may reduce or enhance the value of a received outcome (329). Thus contexts add EU\textsubscript{cxt} to the EU of odors, flavors, soft drinks, and wine and gambles to modify EU\textsubscript{net}.

5. Economic cost

Natural rewards do not only provide nutrients and calories, their acquisition usually requires calorie expenditure that reduces the gains provided by them. We need to work for them, approach them, collect them, and consume them. This intuition applies to all rewards and is conceptualized as economic cost. Every good has its cost, and we usually calculate the net benefit before handing over the money. We need to ensure that the cost does not wipe out the benefit. The value of Eve’s apple is reduced by the loss of paradise. Net benefit utility can be calculated generically as a subtraction

\[
\text{EU}_{\text{net}} = \text{EU} - \text{EU}_{\text{cost}} \quad (13)
\]
Equation 13 is based on physical net benefit, physical income, and physical cost (such as calories)

\[ \text{Net benefit} = \text{income} - \text{cost} \quad (14) \]

and becomes net utility by applying different utility functions \( u(\text{income}) \) and \( d(\text{cost}) \)

\[ \text{Unet} = u(\text{income}) - d(\text{cost}) \quad (15) \]

Income utility \( u(\text{income}) \) is often concave (gradually flattening due to decreasing marginal utility) (FIGURE 28A, top). Physical effort, as the energy expenditure required to obtaining a reward, is an economic cost, as long as it is associated with the pursuit and acquisition of that reward. As higher efforts become increasingly harder, cost is often assumed to be convex (FIGURE 28A, middle). With convex cost, the disutility of high effort may outweigh the increase of income utility, resulting in decreasing Unet and loss of monotonicity (FIGURE 28A, bottom). The effort cost in Equations 13–15 needs to be distinguished from temporal discounting, which may be partly due to another form of cost, opportunity cost, and reduces reward value through a separate mechanism (see Equations 16-18A).

A gross simplification may neglect the different curvatures possible with Equation 15 and apply a common logarithmic utility and disutility function

\[ \text{Unet} = \log(\text{income}) - \log(\text{cost}) \quad (15A) \]

which is equivalent to the benefit-cost ratio

\[ \text{Unet} = \log(\text{income/cost}) \quad (15B) \]

Appropriate behavioral tests would involve separate assessment of income-utility and cost-disutility functions in a common reward value currency, using choices between reward alone [to obtain \( u(\text{reward}) \)] and reward combined with cost [to obtain \( d(\text{cost}) \)], similar to composite outcomes involving punishers (FIGURE 11A).

In analogy to temporal discounting (see Equations 17 and 18 below), alternative effort models used in some neurobiological studies (419, 439) incorporate effort cost into net benefit by exponential discounting

\[ \text{Net benefit} = \text{income} \times e^{-k \times \text{cost}} \quad (15C) \]
to result in a utility measure

\[ \text{Unet} = u(\text{income}) \times e^{-k \times \text{cost}} \]  

(15D)

or by hyperbolic discounting

\[ \text{Net benefit} = \frac{\text{income}}{1 + k \times \text{cost}} \]  

(15E)

\[ \text{Unet} = \frac{u(\text{income})}{1 + k \times \text{cost}} \]  

(15F)

where \( k \) is discounting factor, reflecting subjective cost sensitivity.

The pursuit of any reward reduces the chance of obtaining another reward. This opportunity cost arises with mutually exclusive rewards and is defined as the value of the highest foregone alternative reward. Thus, in principle, Equations 13-15D apply to every reward obtained in the presence of a viable, mutually exclusive alternative. Opportunity cost applies also when rewards are delayed and other beneficial options cannot be pursued, which is conceptualized as temporal discounting (see below).

Movements and effort may not only constitute a cost but also have intrinsic reward value (31, 358, 546, 547). A behavior may be undertaken for the pure pleasure of it and without resulting in any other reward. A kid is having fun building an airplane without getting monetary or social benefits. The intrinsic reward value can be considered as income, the required effort (and missed opportunity for homework) is the cost, and Equation 15 becomes

\[ \text{Unet} = u'(\text{intrinsic income}) - d(\text{cost}) \]  

(15G)

with \( u' \) as utility function distinct from \( u \). Intrinsically rewarding behavior may also lead to extrinsic reward. A mechanic loves to rebuild historic car engines and also sells them well, a typical scenario for people that love their jobs. Here, income utility derives from both intrinsic and extrinsic reward value, and Equation 15 becomes

\[ \text{Unet} = u(\text{extrinsic income}) + u'(\text{intrinsic income}) - d(\text{cost}) \]  

(15H)

Importantly, the added value from intrinsic rewards may outweigh high effort costs. Even if \( u(\text{extrinsic income}) \) in Equation 15H is below \( d(\text{cost}) \), Unet remains positive, and the behavior viable, if the utility gained from intrinsic reward \( [u'(\text{intrinsic income})] \) is sufficiently high. People that love their jobs may be content with unspectacular pay. Alternative formulations may consider intrinsically rewarding behavior as reward in its own right and apply Equation 25 for multiple rewards (see below). Taken together, effortful behavior may both enhance and reduce net benefit. As the intrinsic gains and costs of behavior are subjective and not deductible from physical movement measures alone, it is particularly important to assess the involved utilities in choices against a reference utility using quantitative psycho-

physical methods. Simply assuming that movements represent only an economic cost, and estimating its utility from physical rather than behavioral measures, may be misleading.

Behavioral studies on cost follow primarily Equation 14. Experiments involve rats crossing barriers in T mazes (499) or working for intracranial self-stimulation (62) and monkeys performing sequential movements (60) or exchanging tokens for reward (69). Tests on humans sampling different numbers of targets comply with Equation 15B (108). Binary choices increase with reward value and decrease with effort cost measured as number of lever presses in monkeys (263) and rats (115, 116, 621) and lever movement against resistance in monkeys (233, 419) (Figure 28A). These choice procedures allow comparison of subjective net benefit between the options analogous to Unet, but without explicitly assessing \( u(\text{income}) \) and \( d(\text{cost}) \). Behavioral choices are fit better by the physical cost subtraction model of Equation 14 than by exponential or hyperbolic cost discounting of Equations 15C and 15E (233). Reaction times increase with increasing effort, well fit by exponential or hyperbolic cost discounting of Equations 15C and 15E (419).

Neurons in anterior cingulate cortex show activity increases with income and decreases with effort cost (233, 263), thus following the components of Equation 14. In contrast, orbitofrontal and dorsolateral prefrontal cortex neurons code either income or cost but not both (233). Neurophysiological responses and dopamine release evoked by reward-predicting cues in nucleus accumbens decrease with higher fixed ratio effort (115, 116) (Figure 28B). Exponential or hyperbolic cost discounting (Equations 15C and 15E) affects some dopamine neurons in substantia nigra (419). Thus neuronal responses in reward structures typically reflect the influence of effort cost on reward value.

The dopamine sensitivity to effort is at odds with the failure of finding reductions in voltammetric dopamine reward responses in nucleus accumbens with effort (224, 621). However, the effort in these studies was intercorrelated with temporal discounting, which is well established to reduce reward value and corresponding electrophysiological and voltammetric dopamine responses (116, 158, 285) (Figure 28D). Thus it is unclear why not at least temporal discounting reduced the dopamine responses. The effort-insensitive voltammetric dopamine changes were slow (lasting \( >8 \) s) and one order of magnitude lower than the typical phasic voltammetric dopamine changes \( [0.5-15 \text{ nM} (224, 621) \text{ versus } 30-70 \text{ nM} (116)] \), which raises additional, methodological issues. The slower responses resemble dopamine changes of similar durations and magnitudes that have no correlate in dopamine impulse activity (237). Thus the slower, lower, effort insensitive dopamine changes likely do not reflect subjective value (224) but unlikely bear on the
role of phasic dopamine signals in economic value coding. Thus the slower, lower, effort insensitive dopamine changes may not reflect subjective value (224) which, however, does not bear on the phasic dopamine coding of all subjective value components tested so far, including reward amount (598), reward type (301), risk (156, 301, 560, 570), delay (116, 158, 285), and effort cost (116) all the way to formal economic utility (560).

6. Temporal discounting

Typical primary nutrient rewards, like fruits or meat, mature and decay over time. Once consumable, they risk destruction from many causes including competition by other reward seekers. Rewards often need to be consumed at particular times due of specific energy demands, and later rewards are less tangible in general. Thus specific rewards in specific situations should be harvested at specific moments. This is why we cannot wait for the pay slip, use refrigerators, drink wine at its best age, and mate only after maturation. These temporal constraints favor temporal sensitivity in the evolution of the brain’s reward system and lead to personal, potentially beneficial traits of impatience, impulsivity, and risk taking. The characteristics extend to all rewards and explain the variation of subjective reward value over time. Although temporal discounting likely originates in the perishable nature of goods, it occurs even when the reward remains physically unchanged and overall reward rate is held constant, thus dissociating subjective from objective reward value.

Temporal discounting studies focus mostly on monotonic, forward decreases of value (4, 462, 473). However, value may evolve nonmonotonically in time as birds store food (443), often increasing initially and then decreasing, although the inverse is also possible. The change of reward value over time occurs between a reward-predicting stimulus and reward delivery when no operant action is required (Pavlovian conditioning) or between an action and the resulting reward. In animal learning theory, increasing temporal delays decrease the proximity (contiguity) between stimulus or action and reinforcer. Delays position the reward away from the stimulus or action and more into the background, thus reducing the subjectively perceived reward contingency (dependency of reward on stimulus or action) (175, 220). Thus both subjective devaluation and subjective contingency reduction reduce the power of reinforcers for learning.

Temporal discounting constitutes an opportunity cost when waiting prevents individuals from other activity leading to reward or from using the reward for beneficial investments. Thus temporal discounting can be stated generically as an economic cost, in analogy to Equation 13

\[ E_{\text{Unet}} = EU - E_{\text{Utemp}} \]  

Obtaining net benefit \( E_{\text{Unet}} \) requires two steps, assessment of \( u(m) \) from objective reward magnitudes and then its temporal discounting (296). Temporal discounting can be implemented by exponential functions, resulting in constant discount rate

\[ E_{\text{Unet}} = u(m) e^{-k*D} \]  \hspace{1cm} (17)

where \( k \) is discount factor, \( D \) is subjective delay, as assessed from maximal responses at the subjectively perceived delay (peak interval procedure, Ref. 469). Alternatively, hyperbolic discounting may describe the process better. The generalized hyperbola combines the two models by adding a free parameter \( a \) (325)

\[ E_{\text{Unet}} = u(m)/(1 + k*D)^{a} \]  \hspace{1cm} (18A)

Hyperbolic discounting is particularly prominent against immediate rewards. When all rewards have longer delays, exponential discounting may describe the process better. The generalized hyperbola combines the two models by adding a free parameter \( a \) (325)

\[ E_{\text{Unet}} = u(m)/(1 + k*D)^{a} \]  \hspace{1cm} (18A)

Unet can be assessed psychophysically with intertemporal choices between a variable early and a constant delayed reward (adjusting-amount procedure). Unet of the late reward is inferred from the utility of the early reward that produces choice indifference. Unet in monkeys decreases monotonically across reward delays of 2, 4, 8, and 16 s by ~25, 50, and 75%, respectively (Figure 28D, blue) (285), with slightly better fits by hyperbolic than exponential discount functions. Temporal discounting decreases from rodents to monkeys to humans.

Some theories of temporal discounting distinguish between a prefrontal system controlling immediate, impulsive consumption and a separate valuation system (352, 538), whereas more straightforward accounts assume that a single reward value system directly discounts value according to Equations 16-18A, although the truth might lie in between. Some reward value coding neurons in all reward systems show decreased responses with increasing delays. Responses of monkey dopamine neurons to reward-predicting stimuli decrease monotonically across reward delays of 2–16 s (Figure 28D, red), despite constant physical reward magnitude, thus matching closely the behavioral discounting (compare blue and red) (158, 285). Corresponding to human behavior (277), lower reward magnitudes are associated with steeper neuronal discounting (285). The same dopamine neurons code reward magnitude. Temporal discounting may reflect reduced value at the time of the predictive stimulus (input stage) or during the decision process (output stage). The reduced dopamine response occurs in imperative tasks without choices, suggesting value alterations in the input stage. The reduced neuronal response leads to reduced dopamine release in nucleus accumbens (116). Inversely, dopamine prediction error responses to reward delivery increase with longer delays. Both responses are fit better by hyperbolic than exponential func-
tions. The almost perfect overlap between behavioral and neuronal discounting (FIGURE 28D) suggests that temporal delays affect dopamine responses via changes in subjective reward value. Dopamine neurons process delayed rewards as having less value.

Many reward value coding neurons in other reward structures show temporal discounting, including orbitofrontal cortex (478), prefrontal cortex (272, 475), anterior cingulate (233), premotor cortex (475), parietal cortex (477), and dorsal and ventral striatum (79, 115). Reversal of cue-delay associations leads to reversed neuronal responses, suggesting an influence of delay rather than visual properties (475). Often the same neurons code also reward magnitude (476), suggesting common subjective value coding across different rewards. Taken together, neuronal population in all major reward structures code subjective reward value during temporal discounting.

7. Risk

Rewards are, as all environmental events, inherently uncertain. The uncertainty can be formally approached via the description of rewards as probability distributions of values. In contrast to the first statistical moment of probability distributions that defines value (EV, Equation 8), the higher moments reflect risk, including variance, skewness, and kurtosis (FIGURE 29A). Risk refers to the degree of uncertainty in known probability distributions, whereas ambiguity denotes uncertainty when probability distributions are incompletely known. Risk, and also ambiguity, has at least three functions in reward processing. First, individuals should perceive risk appropriately, irrespective of making a decision at a given moment. Second, as the terms of risk avoidance and risk seeking indicate, risk affects the subjective value of rewards. These two functions, and their processing by reward neurons, will be described in this section.
Third, risk informs about the spread of reward probability distributions which neurons may take into account when processing reward value within an existing distribution. This function will be described subsequently as adaptive processing.

Popular risk measures are variance

\[
\text{var} = \sum [p_i (m_i - \text{EV})^2]; \text{ over all } i \tag{19}
\]

where \(P\) is probability and \(m\) is magnitude. Standard deviation (SD), the square root of variance, is also a good and frequently used risk measure. EV and SD fully define Gaussian and equiprobable binary probability distributions. However, the coefficient of variation, \(CV = SD/EV\), is often a better measure for risk in economic choices (633). Other risk measures are Shannon informational entropy (analogous to thermodynamic entropy) that seems appropriate for neuronal information processing systems (477).

Although probabilistic rewards are intrinsically risky, probability is not a monotonic measure for risk but for value \((Equation \ 8)\). Whereas reward value increases monotonically with probability, variance as a measure for risk follows an inverted U function. Whereas reward value increases monotonically with probability, variance as a measure for risk follows an inverted U function. EV and SD fully define Gaussian and equiprobable binary probability distributions. However, the coefficient of variation, \(CV = SD/EV\), is often a better measure for risk in economic choices (633). Other risk measures are Shannon informational entropy (analogous to thermodynamic entropy) that seems appropriate for neuronal information processing systems (477).

Risk is perceived subjectively through genuine risk avoiding/seeking tendencies (“I hate/love the risk”), risk distortions (“I have no idea/am confident about the outcome”), the value at which loss/win occurs (“no gambling at high values/peanuts”), the domain, situation, and familiarity of the risky event, and hormones (86, 87, 630, 632). Some of these factors may play a role when defining subjective risk in analogy to Equation 19 as variance of utility

\[
\text{varU} = \sum \pi [p_i (m_i) * (u(m_i) - \text{EU})^2]; \text{ over all } i \tag{20}
\]

A fuller account would include more of the mentioned subjective factors.

It is important for the understanding of risk attitude to distinguish between objective and subjective value. The famous example is the hungry bird with challenged energy balance (87). The bird requires 100 calories to survive the night and has the choice between a certain option of 90 calories and an equiprobable gamble of 0 and 110 calories \((P = 0.5)\). The bird would hopefully choose the gamble, as this provides it at least with a 50% chance of survival. However, from the perspective of objective value, the usually risk-avoiding bird seems to be surprisingly risk seeking, as it prefers a risky option with an EV of 55 calories over a certain 90 calories. However, the bird’s behavior follows first-order stochastic dominance when we consider survival values \(0\) vs. \(0/1\) survival, which tests for meaningful and rational choices (341).

Risk avoidance and risk seeking suggest that risk affects economic choices, which can be visualized via the curvature of the utility function \((Equation \ 9)\). With concave utility \((FIGURE \ 29C)\), the initially steeper slope indicates higher marginal utility in lower ranges. Losses move utility into the lower, steeper range and thus appear subjectively more important than gains which move utility into the upper, flatter range, resulting in risk avoidance. In contrast, with a convex utility function \((FIGURE \ 29D)\), the steeper slope in higher ranges makes gains appear subjectively more important than losses, thus encouraging risk seeking. In some cases, the curvature of utility functions can change, often from risk seeking with low values to risk avoidance with high values. Risk seeking at low values may suggest that the attraction of gambling outweighs potential losses (“peanuts effect”) (152, 436, 630).

The subjective value of risk combines with riskless utility to result in the utility of risky choices, as expressed generically

\[
\text{EU}_{\text{net}} = \text{EU}_{\text{r}} - \text{EU}_{\text{risk}} \tag{21}
\]

with \(\text{EU}_{\text{r}}\) as riskless utility, and \(\text{EU}_{\text{risk}}\) as disutility due to risk. This approach recognizes the difference between riskless and risky utility functions (631). Positive \(\text{EU}_{\text{risk}}\) reduces \(\text{EU}_{\text{net}}\) and characterizes risk avoidance, whereas negative \(\text{EU}_{\text{risk}}\) enhances utility and characterizes risk seeking. This approach follows in kind the construction of utility from objective measures of risky outcomes \(\text{EV}\) and variance) by the mean-variance approach in financial economics based on Taylor series expansion (238, 315)

\[
\text{EU} = \text{EV} - b * \text{variance} \tag{21A}
\]

with \(b\) as subjective risk weighting coefficient. Equation 21A applies, strictly speaking, only to quadratic utility functions.

In addition to subjective risk perceptions, the subjective effect of risk is determined by the curvature of utility functions that are typically concave \((FIGURE \ 29C)\) and asymmetric for gains and losses (steeper gain functions, or steeper loss functions inducing “loss aversion”) (258). Subjective pessimistic/optimistic probability distortions are further subjective influences (“today is my bad/lucky day”). These subjective factors transform symmetric variance risk into the common asymmetric notion of risk as the danger to lose (higher loss than gain utility) \((FIGURE \ 29E)\).

The most simple and controlled risky outcome is provided by a binary, equiprobable gamble \((P = 0.5)\) each outcome in which risk can be varied without changing the EV (mean-preserving spread) and which has negligible, constant probability distortion and is not skewed (497) \((FIGURE \ 30A)\). Risk attitude can be estimated by eliciting the certainty
Above 0.5–0.8 ml (CE liquid volumes (CE)), or sometimes as EV/CE (1 for risk avoiders, 0 for risk seekers, 0 for risk neutrality), or sometimes as EV/CE (1 for risk avoiders, 0 for risk seeking, 0 for risk neutrality) (558). Rhebus monkeys are often risk seeking with standard small liquid volumes (CE > EV) and risk avoiding with volumes above 0.5–0.8 ml (CE < EV) (FIGURE 29B) (355, 560, 644). Their utility function is initially convex, then linear and ultimately concave (FIGURES 24D AND 27A), thus showing an inflection as conceptualized and found in human utility functions (152, 339, 436, 630).

Formal tests of stochastic dominance (341) confirm that monkeys understand the risky stimuli (560). First-order stochastic dominance helps to define meaningful choices without requiring assessment of a utility function. However, higher EV does not necessarily satisfy first-order stochastic dominance, due to non-linear utility, non-linear probability weighting, and risk attitude. In statewise dominance, a case of first-order stochastic dominance, a gamble dominates another gamble if all its outcomes are equal except at least one outcome that is better. Monkeys’ behavior satisfies this criterion in choices between a certain reward and an equiprobable, binary gamble whose low or high outcome equals the certain reward. They avoid the low certain reward (set at low gamble outcome, left) and prefer the high certain reward (set at high gamble outcome, right). Averaged data from binary choices involving four gambles: 0.1–0.4 ml, 0.5–0.8 ml, 0.9–1.2 ml, 0.1–1.2 ml. D: choices for riskier options with mean preserving spread satisfy second-order stochastic dominance for risk seeking in monkeys, indicating meaningful incorporation of risk into choices. When presented with gambles in the risk-seeking domain of the utility function (left; expected utility, EU, higher for riskier gamble, red), monkeys prefer the riskier over the less risky gamble. (B–D from Stauffer et al. (550)).

The orbitofrontal cortex plays a key role in risky choices (317). Human patients and animals with lesions in orbitofrontal cortex or nucleus accumbens show altered risk and ambiguity sensitivity distinct from value (35, 88, 174, 240, 268, 364, 370, 445). Correspondingly, specific orbitofrontal neurons in monkeys show responses to stimuli predicting binary, equiprobable gambles (FIGURE 31A) (391). These risk responses occur irrespective of visual stimulus properties and only very rarely reflect reward value, which is coded in separate orbitofrontal neurons. The absence of value coding argues also against the coding of salience, which is associated with both risk and value. Some monkey orbitofrontal neurons are activated by unsigned (absolute)
FIGURE 31. Neuronal risk processing in monkeys. 
A: coding of risk in single neuron of orbitofrontal cortex. 
Height of bars indicates liquid reward volume, two bars within a rectangle indicate an equiprobable gamble ($P = 0.5$ each). The three gambles have same mean but different variance (9, 36, and 144 ml $\times 10^{-4}$) (mean-preserving spread). This neuron, as most other orbitofrontal risk neurons, failed to code reward value (not shown). [From O’Neill and Schultz (391), with permission from Elsevier.] B: coding of risk prediction error in orbitofrontal cortex. Top: risk prediction error (colored double arrows), defined as difference between current risk (vertical distance between bars in each gamble indicates standard deviation) and predicted risk (common dotted red line, mean of standard deviations of the three gambles). Colored double vertical arrows indicate unsigned (absolute) prediction errors in variance risk. Bottom: averaged population responses from 15 orbitofrontal neurons to unsigned risk error. Such signals may serve for updating risk information. [From O’Neill and Schultz (392).] C: risk coding in single dopamine neuron during the late stimulus-reward interval. The activation is maximal with reward probability of $P = 0.5$, thus following the inverted U function of variance risk shown in FIGURE 29B. Trials are off line sorted according to reward probability. [From Fiorillo et al. (161).] D: independence of risk activation in single dopamine neuron from reward outcome in preceding trial. Activation is not stronger after positive prediction errors (top) than after negative prediction errors (bottom). Thus prediction error responses do not backpropagate in small steps from reward to stimulus. Both rewarded and unrewarded current trials are shown in each graph. [From Fiorillo et al. (162).] E: influence of risk on subjective value coding in dopamine neurons. The population response is stronger with subjectively higher valued certain (safe) juice rewards (green: blackcurrant; blue: orange; both 0.45 ml), demonstrating value coding. This value response is enhanced with risky juice volumes (equiprobable gamble of 0.3 and 0.6 ml, same mean of 0.45 ml), corresponding to the animal’s risk seeking attitude (higher subjective value for risky compared with certain rewards, inset). [From Lak et al. (301).] F: dopamine responses satisfy first-order stochastic dominance, demonstrating meaningful processing of stimuli and reward value under risk (averages from 52 neurons). In the two gambles (blue, red), the higher outcomes are equal, but the lower red outcome is higher than the lower blue outcome, defining gamble dominance. G: dopamine responses satisfy second-order stochastic dominance, demonstrating meaningful incorporation of risk into neuronal utility signal (averages from 52 neurons). The red gamble has greater mean-preserving spread than the blue gamble, defining dominance with risk seeking. [F and G from Stauffer et al. (560).]
risk prediction errors, defined as unsigned difference between current and predicted risk, which may serve for updating risk information [FIGURE 31C] (392). In a categorical discrimination task, responses in orbitofrontal neurons reflect the riskiness of olfactory stimuli but in addition differentiate between correct and erroneous choices in correlation with decision confidence (264). The orbitofrontal risk signal is distinct from movement-related activity (153), some of which correlates with particular, asymmetric forms of salience (393). In human prefrontal cortex, BOLD signals show direct risk coding, distinct from value (599), which reflects risk perception irrespective of choice. Outside of orbitofrontal cortex, supplementary eye field neurons differentiate categorically between certain and risky trials and code certain and risky reward value differentially (551), and septum neurons show slow risk responses closely resembling the risk ramp of dopamine neurons (373).

Dopamine neurons code reward risk during the stimulus-reward interval after the value prediction error response [FIGURE 31C] (161). The activation ramps up to the reward and varies with probability in an inverted U function, analogous to risk measures such as variance, standard deviation, and entropy. It increases with variance and standard deviation, even when entropy is kept constant. Thus the signal reflects risk rather than value [FIGURE 29B]. An analogous signal is also seen in the human ventral striatum, which is distinct from value and most likely reflects synaptic activity induced by the electrophysiologically measured dopamine risk signal (437). Although its latency is too long for immediate decision processes, the dopamine risk signal might affect the subsequent prediction error signal at the time of the reward. In coding risk, the signal varies with surprise salience and thus may affect the learning constant of attentional (associability) learning theories (425, 336). Also, the slower risk signal might contribute to the risk sensitivity of the faster, phasic dopamine signal (see below) (156, 301, 560, 570).

Straightforward implementations of TD models may produce similar ramps by backpropagation of prediction error responses from reward to the preceding conditioned stimulus in small steps via imaginary stimuli (374, 524), thus resembling the dopamine risk response (388). As the backpropagating activation would be small and dispersed, it would appear as ramp only when averaging across multiple trials (183, 388). However, the dopamine risk signal occurs in single trials without averaging and does not backpropagate via imaginary stimuli, as suggested by its absence after prediction errors [FIGURE 31D] (161, 162), but jump in single steps from reward back to the preceding actual stimulus [FIGURE 14A] (117, 410, 598). The ramp is not a necessary feature of TD models but occurs only in specific implementations (374, 388, 524). Biologically more plausible TD implementations and spectral timing models use stimulus eligibility traces, as originally suggested [FIGURE 14B] (574), and model well the single step backward transition of dopamine prediction error responses from reward to the next preceding conditioned stimulus without error backpropagation and thus without a ramp during the stimulus-reward interval [FIGURE 14, C AND D] (71, 410, 573). In particular, models using a biologically consistent striatal and midbrain circuit replicate well the dopamine risk ramp (581). Thus the dopamine ramp represents a genuine risk response and not the necessary byproduct of TD models.

In addition to the late and slow dopamine risk signal, risk affects the phasic dopamine value response. The subjective value reward signal of dopamine neurons is enhanced when monkeys prefer risky over certain outcomes of low or intermediate reward magnitudes [FIGURE 31E] (156, 301), suggesting incorporation of risk into the final construction of utility in possible compliance with Equation 21A. Correspondingly, dopamine concentration changes following risky cues are reduced in risk avoiding rats and enhanced in risk seekers (570). In formal tests informed by economic theory (341, 497), dopamine responses follow first- and second-order stochastic dominance [FIGURE 31, F AND G], suggesting meaningful coding of economic value and risk and incorporation of risk into neuronal substrates of subjective value. Thus three types of experiment demonstrate that dopamine neurons incorporate risk appropriately into neuronal substrates of subjective value and formal utility, namely, value responses modified by risk in correspondence with risk attitude [FIGURE 31E], second-order stochastic dominance between different risky outcomes [FIGURE 310], and processing of utility assessed via CEs across risk seeking and risk avoiding domains [FIGURE 27B] (560).

Risk influences also subjective value coding in orbitofrontal cortex (444, 482) and modifies saccadic activity in posterior cingulate cortex (355). In human prefrontal cortex, risk affects value responses in close correspondence to individual risk attitudes in analogy to Equation 21A (596). Direct risk coding for risk perception and incorporation of risk into subjective value constitute two different processes that allow multiple uses of risk information for behavior.

Taken together, the distinct neuronal coding of value and risk suggest that Pavlovian reward predictions concern probability distributions defined by value and risk. In contrast to reward magnitudes that impact directly on sensory receptors, the coding of EV, probability, and risk of reward is not explained by peripheral sensory stimulation but requires central neuronal mechanisms including memory. Irrespective of these implementation details, the separation of reward signals according to the statistical moments of probability distributions, and the possible correspondence to Taylor series expansion, provides a biological basis for the mathematical treatment of rewards as probability distributions, demonstrates the biological plausibility and imple-
moments of the current probability distribution of objective (EUadapt) that summarily covers adaption to all statistical The consequences of adaptive processing are changes in net discrimination with steepest possible response slopes while gain to loss for identical objective values.

For optimal response slopes and thus best use of the limited neuronal coding range, reward processing may adapt to all sensory intensities (148, 235, 307).

For optimal response slopes and thus best use of the limited neuronal coding range, reward processing may adapt to all statistical moments of reward distributions, including EV and variance (FIGURE 29A). When schematically separating these two moments, changes in EV lead to processing shifts with maintained response slopes (FIGURE 32A), whereas variance affects response slopes, which become steeper with narrower distributions and flatter with wider distributions (FIGURE 32B). These adaptations lead to optimal reward discrimination with steepest possible response slopes while maintaining as wide reward coverage as possible.

The consequences of adaptive processing are changes in net utility, which may be expressed generically by adding a term (EUadapt) that summarily covers adaption to all statistical moments of the current probability distribution of objective values

This approach includes reference-dependent valuation in prospect theory derived from Equation 10

$$EU = \sum_{i} (u(m_i) - r) \cdot p_i(m_i); \text{ over all } i$$

with $r$ as reference value (258). The term $r$ is replaced by EV if reference $r$ derives from a probability distribution rather than being a single value

$$EU = \sum_{i} (u(m_i) - EVr) \cdot p_i(m_i); \text{ over all } i$$

Equation 22B describes adaptation to the first moment of the probability distribution of objective values. The adaptation to EV can lead to vastly different utilities for identical objective values (FIGURE 32A).

A milder approach to reference-dependent valuation combines nonadapted and adapted utility (290). This approach would refine the generic Equation 22 to

$$EUnet = \alpha EU + \beta (EU - EUadapt); \alpha + \beta = 1$$

$\alpha$ and $\beta$ are additional weighting coefficients. In the terminology of Köszegi and Rabin (290), EU represents consumption utility, and (EU-EUadapt) is reference dependent gain-loss utility. Thus EUnet consists of weighted absolute and relative utilities. Considering nonadapted value recognizes the fact that we know cognitively the absolute reward value, even though we value it relative to references or discount it over time. We know that $100$ is still the same $100$ irrespective of having received $50$ or $500$ the last time around, even though we may feel more elated receiving them as a positive than negative surprise.

Adaptive processing may not be limited to a single reference value and should be extended to full distributions. Inclusion of other statistical moments such as variance adds slope changes. We extend Equations 10 and 22B by introducing $\omega$ for slope adaptation to variance (or standard deviation)

$$EU = \sum_{i} \{u(\omega * (m_i - EVr)) \cdot p_i(m_i)\}; \text{ over all } i$$

A convenient term for $\omega$ is 1/standard deviation (384, 438). The adaptation to variance can lead to larger utilities for identical or even lower objective values (FIGURE 32B).

Adaptations to reward probability distributions underlie several interesting behavioral phenomena. The negative contrast effect derives from a downshift of reward and reduces behavioral reactions and preferences compared with the same low reward occurring without preceding higher values (51, 104, 593), even when the lower outcome adds value (513). A positive but less prominent contrast effect exists also. The opponent process theory of motivation ac-
counts for observations indicating that cessation of reward is aversive, whereas end of punishment is rewarding (552). This effect may well derive from short-term adaptation to the initial reward or punishment, respectively. Adaptations to distributions may also explain the dichotomy in humans between “experienced utility” reflecting utilitarian, hedonic value (39) and “decision utility” measured in overt choices, although overweighting of recent and maximal value is also important (259). This effect is also seen in monkeys (52). The distinction may explain irrational, suboptimal, dominated choices. Monkeys favor [one food morsel plus a 50% chance of a second morsel] over [two food morsels with 50% chance of losing one], although EVs are both 1.5 morsels, although endowment effect and loss aversion may also play a role (91).

Reward responses in dopamine neurons, striatum, amygdala, parietal cortex and orbital, dorsolateral, and cingulate prefrontal cortex adapt to reward magnitude and reward probability distributions as suggested by Equation 22 (137, 628). More formally, the adaptations may occur to changes in EV (FIGURE 33A) (42, 106, 232, 601), variance (FIGURE 33B) (284, 420, 598), or both combined (80, 403). In subtracting reward value from the mean (prediction error) and dividing it by standard deviation, dopamine neurons code a z-score of reward value (598). Divisive normalization from sensory physiology describes well the adaptations to changes in individual values of choice options in reward neurons of parietal cortex (332). The ventral striatum is importantly involved, as lesions here reduce reward contrast in rats (326). Taken together, the adaptations would be driven by neuronal EV and risk signals and result in a match between the probability distribution of neuronal responses and the probability distribution of current reward values. The result would be dynamically varying, optimal reward response slopes and thus efficient reward processing for choices. The loss of the actual values would be immaterial in choices comparing the options relative to each other. Thus, in adapting to the two key moments of reward probability distributions, reward neurons make ef-
efficient use of processing resources for optimal reward discrimination and choice.

The EU shift between environments and the slope adaptation may result in lower utility of higher absolute reward values (FIGURE 32B). Thus adaptation may lead to violation of rank order, menu invariance, independence from irrelevant alternatives, and transitivity, which are crucial for meaningful economic choices. However, at least two processes may serve to maintain transitivity in the presence of adaptive valuation. First, the mentioned predictive EU and risk signals in dopamine neurons (598) and orbitofrontal cortex (391) convey information about the two key moments of the reward probability distribution of each environment. These signals may combine with the adapted value signal to generate an accurate, absolute, nonadapted value signal and thus maintain transitivity at the neuronal level. Second, adaptations require time and may only concern subpopulations of reward neurons. Indeed, only a fraction of orbitofrontal reward neurons show adaptation, whereas the remainder codes reward magnitude irrespective of other rewards (403, 406, 284). Although longer test periods might have increased the number of adapting neurons, some orbitofrontal reward neurons seem to code nonadapted, absolute value necessary for maintaining transitivity. Adaptation and absolute value coding by different neuronal populations would combine efficacy with transitivity.

9. Social rewards

Observation of social partners and comparison of rewards between them is crucial for competition and cooperation that improve performance and give individuals access to otherwise unobtainable resources and ultimately enhance fitness. Social factors and contexts have at least two distinct reward functions.

First, viewing and encountering others is a reward on its own. Rhesus monkeys sacrifice quantities of liquid reward for viewing body parts of conspecifics, demonstrating that social images have reward value (124). Neurons in monkey parietal cortex code the reward value of conspecific’s faces and hindquarters (279), although orbitofrontal and striatal neurons code these social stimuli often separately from liquid rewards (629).

Second, reward to others affects own reward processing. With benevolent exceptions (“Mother Teresa” and many other mothers), individuals hate to receive less reward than others when all other factors are equal. Disadvantageous inequity aversion typically induces a sharp, gradually declining, drop in the utility of own reward (FIGURE 34A) (324). Emotional correlates are envy and distaste for unequal or unfair outcomes. Individuals often also dislike, usually to a lesser extent, receiving more reward than others. This advantageous inequity aversion may elicit a similar

---

**FIGURE 34.** Social reward processing. A: social reward inequity aversion. Difference between own and other’s reward reduces my utility. Usually, disadvantageous inequity (getting less than others, negative difference, left) has stronger and more consistent effects than advantageous inequity (getting more than others, right). [From Loewenstein et al. (324)]. B: action-dependent coding of social reward in single striatal neurons. In the imperative reward giving task (modified dictator game), two monkeys sit opposite each other across a touch-sensitive computer monitor and give reward to itself and the conspecific. Left: neuronal activation when receiving own reward, either only to itself (red) or together with conspecific (green), but no activation with reward only to conspecific (violet) or to nobody (blue). Right: activation with own reward occurs only with own action (top, continuous lines) or only with conspecific’s action (bottom, dotted lines, different neurons between top and bottom). [From Báez-Mendoza et al. (25).]
but weaker utility drop. Associated emotions are compassion, sense of fairness, guilt, dislike of inequity, positive reciprocity, group welfare, “warm glow,” and maintaining personal social image (83, 149). The aversion may turn into advantageous inequity seeking (getting more than others) with competition and other advantage seeking situations associated with pride, sense of achievement, superiority feelings, and retaliation. Good behavioral tools for assessing inequity are the ultimatum game in which rejection of unfair offers assesses disadvantageous inequity aversion, and the dictator game in which the reward fraction handed to the opponent reflects advantageous inequity aversion (83, 149).

Final own expected utility in social settings EU_{net} can be stated generically as

$$ EU_{net} = EU - EU_{social} \quad (23) $$

EU stands not only for income utility (Equation 10) but also for any utility EU_{net} derived from reward components (Equations 11, 12, 13, 18, 21, and 22). EU_{social} is derived from the difference between own value $x$ and value $x_o$ received by the opponent(s), weighted by disadvantageous ($\alpha$) and advantageous ($\beta$) inequity coefficients in analogy to a popular model (150), resulting in

$$ EU_{net} = EU - \alpha \cdot \max[(x_o - x, 0] - \beta \cdot \max[(x - x_o), 0] \quad (23A) $$

where $\alpha$ and $\beta$ are disadvantageous and advantageous inequity coefficients, respectively. Alternative accounts focus on reciprocal “fairness weights” which enhance own reward value when being nice to a nice opponent and reduce own value when being mean to a mean opponent, thus stabilizing cooperation and disfavoring unsocial behavior (442). Both models suggest that reward to others has appreciable value for the observer. Although condensing social reward value into a common value variable is tempting and would help to conceptualize decisions, nonvalue factors such as strategy may importantly influence decisions (50, 150, 151).

Chimpanzees and rhesus, cebus, and marmoset monkeys show prosocial tendencies and actively observe rewards in conspecifics (77, 123, 359). Cebus show disadvantageous but no advantageous inequity aversion (69). Rhesus prefer receiving own reward alone over both animals receiving the same reward, suggesting advantageous inequity seeking (negative $\beta$ of Equation 23), and they prefer seeing conspecifics receive reward over nobody receiving reward, suggesting disadvantageous inequity seeking (negative $\alpha$) in this setting (23, 89). Disadvantageous inequity reduces the effort cebus are willing to spend (612). Many of these social effects may depend on the relative social positions of the monkeys within their group and would merit further investigation. These social tendencies seem simplistic compared with those of humans but may constitute basic building blocks for social encounters and exchanges.

Reward neurons in monkey orbitofrontal cortex and striatum process primarily own rewards in social settings, whereas neurons in anterior cingulate cortex distinguish between own and other’s reward or sense primarily conspecific’s reward (Figure 34B, left) (23, 25, 89). Striatal reward neurons distinguish between the social agents whose action leads to own reward (Figure 34B, right) (25). These neurons are only active when either own action or the other’s action results in own reward, and many of them do not show this activity with a computer instead of a monkey opponent. Neurons in medial frontal cortex show activations during observation of behavioral errors of conspecifics, sometimes together with observation of reward omission (647). Some premotor mirror neurons distinguish between reward and no reward irrespective of who is receiving it (78). These neuronal signals mediate the distinction between own and others’ rewards and actions and thus convey basic components of social reward function. When testing social competition, reward neurons in prefrontal cortex differentiate between competitive and noncompetitive video game performance (234). In humans, disadvantageous inequity reduces, and advantageous inequity activates, ventral striatal reward responses (165), reflecting $\alpha$ and $\beta$ of Equation 23A (the coding direction may reflect different signs of the two coefficients or different neuronal coding slopes). Taken together, social reward neurons code fundamental components of competition and cooperation.

10. Multiple reward attributes

Single rewards come seldom with a single attribute. Attributes include income utility, cost, delay, risk, and social interaction quantified by Equations 9-23A. Their aggregate subjective reward value can be condensed into a single variable called total utility. This has advantages. First, extracting the variable once from the multiple reward attributes and then using it for choices would save the decision maker computing time in crucial moments and thus enhance evolutionary fitness, even if occasional attribute changes require partial recalculations. Second, the brain could transform the condensed neuronal value responses into signals for specific decision variables for competitive decision mechanisms such as winner-take-all (see below).

Total utility $EU_{total}$ of a single reward with multiple attributes can be derived from a combination of attributes, each with their own expected utility derived from likely nonlinear and possibly nonmonotonic utility functions

$$ EU_{total} = EU + f_1 \cdot EU_{state} + f_2 \cdot EU_{ext} - f_3 \cdot EU_{cost} - f_4 \cdot EU_{discount} - f_5 \cdot EU_{risk} - f_6 \cdot EU_{ref} - f_7 \cdot EU_{social} + \text{INTER} \quad (24) $$
where $f_i$ to $f_r$ are weighting coefficients reflecting the different contributions of each reward attribute. In some cases, the attributes can be so closely linked to individual rewards that the total utility $u(\text{total})$ of a reward can be directly derived from adding or subtracting their utility $u(\text{att})$ to the income utility $u(m)$

$$u(\text{total}) = u(m) - u(\text{att})$$ (24A)

Establishing a full probability distribution from $u(\text{total})$ of all rewards would allow computation of EU with Equations 9 or 10 that amounts to the same EUtotal as stated in Equation 24.

Take the example of a hamburger in a pub and apply Equations 10, 11, 12, 13, 18, 21, 22, and 23. Its expected utility EU is defined by the probability distribution of the utilities of the slightly varying hamburger sizes $u(\text{burger})$. My hunger enhances EU by EUstate, and the familiar pub context associated with food and drink adds to the burger’s attraction with EUctx. However, the burger does not come for free (EUcost) and takes a moment to cook because of a busy kitchen, subjecting it to temporal discounting (EUdiscount). Burger quality varies, to which I am averse, thus reducing its EU by EURisk, although it is still the best food in the pub, which puts its EU above a mean reference and thus lose EURef. When the burger finally arrives, it is smaller than my neighbor’s and thus elicits disadvantageous inequity aversion, but better than my dieting other neighbor’s salad, inducing mild advantageous inequity aversion, which both reduce the burger’s EU by EUsocial. Some of these attributes are known to interact which is captured summarily by INTER. For example, smaller rewards are discounted steeper (lower EUdiscount with smaller EU) (277) and larger stakes increase risk aversion (lower EURisk with larger EU) (152, 630).

11. Multiple rewards

Most rewards are composed of several individual reward objects whose values add up and may interact. Our hamburger meal is composed of the burger with its calories and taste, a tomato salad with taste and appealing visual aspects, and a pint of ale with water, calories, taste, and alcohol, thus totaling eight individual rewards. However, ale is sometimes poorly pulled (not in my pub), which would constitute an aversive component that would need to be subtracted from the positive components to obtain a summed final reward. If utilities of several rewards are assessed separately on a common scale, they can be weighted and added up to a summed utility, with possible interaction (INTER)

$$\text{EUsum} = \sum_i (f_i \cdot \text{EUtotal}_i) + \text{INTER}; \quad i = 1, n \text{ rewards}$$ (25)

with $f_i$ as weighting coefficients. Each of the rewards is likely to have its own saturation dynamics. INTER models multisensory integration as interaction that can make aversive taste or odor in some, usually small quantities add to reward value, rather than subtract from it, like pepper on meat, spicy mustard on sausage, or quinine in sugary fizzy drinks.

Equation 25 can be used for assessing the value of an otherwise unquantifiable reward on a scale defined by a common reference reward. For example, in choices between [juice plus picture] versus [juice alone], the value of the picture is reflected by the sacrificed amount of its associated juice to obtain choice indifference (124). Punishers can be included into Equation 25 where they reduce EUsum. They constitute negative value but not economic cost, as they do not reflect energetic, temporal, or monetary expenditures. Behavioral choices of monkeys reveal the value summed from rewards and punishers compatible with Equation 25 (FIGURE 11A) (160).

With the specification of final utility EUsum, the reinforcement Equations 10A–D can be restated as

$$\text{UPESum}(t) = \lambda(t) - \text{EUsum}(t)$$ (26)

$$\text{EUsum}(t + 1) = \text{EUsum}(t) + \alpha \cdot \text{UPESum}(t)$$ (26A)

$$\text{UTDPEsum}(t) = \left[\lambda(t) + \gamma \cdot \text{EUsum}(t)\right] - \text{EU}(t - 1)$$ (26B)

$$\text{EUsum}(t + 1) = \text{EUsum}(t) + \alpha \cdot \text{UTDPEsum}(t)$$ (26C)

Analogous definitions can be made for utility prediction errors of multiple attributes of single rewards by replacing EUsum by EUtotal of Equation 24. With multiple rewards and reward attributes, summed, single prediction errors of Equations 26 and 26B would not require separate computations of utility prediction error for every reward and attribute. Such prediction errors can be used by economic decision mechanisms for straightforward EU updating by standard reinforcement rules.

Alternatively, prediction errors may be computed from separate, possibly multi-attribute, rewards that are added up in a weighted manner, in analogy to Equation 25

$$\text{UPESum}(t) = \sum_i [f_i \cdot \text{UPEtotal}_i(t)] + \text{INTER}; \quad i = 1, n \text{ rewards}$$ (26D)

$$\text{UTDPEsum}(t) = \sum_i [f_i \cdot \text{UTDPEtotal}_i(t)] + \text{INTER}; \quad i = 1, n \text{ rewards}$$ (26E)

A similar calculation is used for incorporating fictive outcomes into gamble values (329). This computation takes the prediction error of each reward into account. An analogous breakdown of prediction error computation, and hence updating, can be made for individual reward components listed in Equation 24.
Given that the dopamine prediction error signal codes utility (560), the dopamine responses (Equations 10E and 10H) can be restated for multiple rewards by replacing $EU$ by $EU_{sum}$

$$DaResp(t) \sim \lambda(t) - EU_{sum}(t) \quad (26F)$$

and for the temporal difference error

$$DaResp(t) \sim [\lambda(t) + \gamma EU_{sum}(t)] - EU_{sum}(t-1) \quad (26G)$$

$DaResp$ in Equations 26F and 26G would apply also to utility error signals in other brain structures.

Common scale value coding is a prerequisite for applying Equation 25 to neuronal signals and is seen in orbitofrontal cortex with different liquid rewards (405) and images of body parts (279), and in dopamine neurons with different liquid and food rewards (301) (FIGURE 23, B AND D). Neuronal signals reflect the sum of positive and negative values from liquid or food rewards and aversive liquids or air puff punisher in monkey dopamine neurons (FIGURE 11D) (157, 160) and anterior cingulate cortex (10), and from odor rewards in human orbitofrontal cortex (190). Compatible with coding negative value as component for Equation 25, some frontal cortical neurons are depressed by losses of gained rewards (532). Ideally, neuronal investigations would test $EU_{sum}$ with all constituent components, although this is impractical and experimenters usually investigate only a few of them at a time.

D. Economic Decision Mechanisms

1. Component processes of decisions

In a highly simplified scheme, economic decisions may involve five component processes (FIGURE 35). The initial intentional component comprises the desire of obtaining a reward (left). This emotion is based on the knowledge, or at least a hunch, that there is a reward that is worthwhile to pursue. With a belief that an action can be attributed to the reward, an intention can be formed to perform an action to get the reward. Thus the reward becomes the goal of the intentional action. In the practicalities of laboratory experiments, these processes can be driven by explicit, temporally specific stimuli indicating the availability of rewards or arise spontaneously by processes internal to the decision maker based on environmental information. The desire to get a
beer may be driven externally by the sight of a pub or internally by the heat of a summer day.

Once the intentional stage has been completed, the inputs to the decision mechanism are processed (FIGURE 35, top). Perceptual decisions require stimulus detection and then evidence accumulation for identifying the sensory reward component, tracking the object and assessing its associated salience components (FIGURE 1, left and middle). Correct identification of each reward option is the basis for choosing the best one. For example, perceptual decisions identify motion direction from the degree of dot motion coherence; the percentage of correct identification translates into reward probability (134, 389).

The perceptual stage is followed by valuation of the individual reward objects (FIGURE 1, right) or, depending on the characteristics of the decision, by valuation of the actions required to obtain each reward object. Valuation comprises the immediate identification or gradual accumulation of value evidence, similar to the accumulation of sensory evidence in perceptual decisions. The process may either parallel the gradual identification of the object or action or start only after they have been completely identified. The decision stage (FIGURE 35, middle) begins with object and action valuation and comprises the selection of the best reward by comparing instantaneously or gradually the values of the available objects or actions. Valuation and value comparison may evolve partly in parallel (double pink arrows). Purely sensory and motor aspects no longer play a role, unless they affect value, such as effort cost (Equation 13). Following valuation and value comparison, the abstract decision specifies the result of the process. It concerns only the decision itself and leads to three selections: the winning object, the winning action, and the prediction of the reward value extracted from the winning object values and action values. If the decision primarily specifies the winning object, subsequent object-action mapping determines the action required to obtain the reward object.

Obtaining the selected object and performing the selected action requires motor decisions for preparing and executing the necessary action (FIGURE 35, bottom), as every decision will ultimately lead to an action for acquiring the chosen reward. The quality of motor decisions determines the efficacy of the economic decision process. Decision makers want to get the action right to get the best reward. In cases in which different possible actions have the same cost and lead to the same reward, motor decisions may become independent of economic decisions. In these cases, the choice of the particular action may depend on noise or on motor exploration. Apart from equivalement actions, it is unclear how much motor decisions can be independent of economic consequences. If actions are chosen for their own sake, they have intrinsic reward value and thus are economic decisions. After experiencing the selected reward object, its value is compared with the value predicted during the decision process (FIGURE 35, right). The resulting prediction error is conveyed to the valuation process in the input stage where it serves to update object values and action values.

The scheme of FIGURE 35 should not suggest strictly sequential processing of decision variables that transitions from sensory identification via value assessment to decision-making. Real life situations are often characterized by noisy evidence, unstable values, moving objects, and changing actions and thus are likely to engage several of these processes in parallel. Furthermore, dopamine, striatum, amygdala, frontal, and parietal neurons show not only hugely different but also considerably overlapping reward valuation and decision activity. Several arguments suggest nonsequential processing of evidence, value, and decisions. First, dorsolateral frontal and posterior parietal neurons are involved in spatial movement processes but also show considerable reward value coding (FIGURES 20E AND 21) (313, 381, 433, 627). Second, ramping activity for evidence accumulation is the hallmark of cortical perceptual decisions but occurs also in subcortical structures during reward expectation and valuation (18, 40, 215, 221, 260, 523). Third, abstract decision coding and chosen action coding overlaps with reward valuation in subcortical structures, including amygdala (188) and striatum (79, 275, 306). Although the different processes may involve different neuronal pools within the same brain structures, the data nevertheless question a strictly sequential processing scheme and suggest more interactive and overlapping processing of evidence, value, and decision-making within the different structures and through subcortical loops.

In addition to many nonvalue factors that determine economic choices (see above), even purely value-based decisions may not involve the complete or correct consideration, computation, or comparison of values. A first important deviation from proper value processing is due to noise. Noise can be derived from uncertain sensory evidence (experimentally modeled, for example, by partly coherent random motion), varying salience attribution, imprecise estimation of economic value (economic noise, for example, when estimating probabilities from changing frequencies, or when deriving value from these probabilities), and noisy learning constants inducing unstable value predictions. Except for sensory evidence, which is external to the brain, the noise results from neuronal noise in sensory receptors and subsequently engaged neurons. Ways to deal with noisy information are averaging out the noise and accumulating the evidence over time. A second deviation from proper value processing occurs with exploration. Options that were previously suboptimal might have improved in the meantime and should better be recognized to prevent missing the best options and ensure utility maximization in the long run. Exploration is captured by $e$-greedy and softmax functions (575). A single parameter models the proportion
of exploration and dominated choices (higher ε and softmax “temperature” for more exploration). Thus utility maximization involves a healthy mix of exploitation (choosing known best options) and exploration (choice of dominated but potentially better options). The proportion of exploration should be fine tuned to the dynamics of the environment to result in maximal reward. It should be higher in volatile (higher ε or temperature) compared with more stable situations (lower ε or temperature). A third deviation from proper value processing occurs with satisfying when individuals fail to consider all available options and choose only from a subset of many options that seem to satisfy their immediate desires (85, 545). Individuals may stop searching for higher value and thus make suboptimal choices. However, in some cases, the cost of considering all options may be high and reduce substantially the net benefit (Equations 13–15), in which cases satisfying may constitute the optimal behavior. A fourth deviation from proper value processing may be due to adaptive value processing that includes inaccessible and thus irrelevant options. In these cases, irrelevant alternatives do affect option values and may make less valuable options appear more valuable and direct choices towards them. The common result of the various forms of incomplete value consideration may be selection of suboptimal, dominated options whose values are lower than those of the best option, a behavior that compromises utility maximization.

2. Decision variables

A decision variable is what the decision maker tries to control and use to achieve the goal. It is the crucial independent variable driving the decision process. The basic decision variable in economic choices is subjective value and, more specifically, formal economic utility. The basic assumption in all economic decision theories is that decision makers aim to maximize utility. Utility is derived from several attributes including reward magnitude, probability, motivational state, context, cost, delay, risk, reference, and social interaction, as stated in Equations 9–25 above. This notion of decision variable differs from the more technical definition that includes all sources of inputs to decision rules and mechanisms (185).

An additional definition of decision variables derives from the specific decision mechanism that serves to control them. Decision variables should be appropriate for this mechanism. Economic decision variables constitute the inputs to the decision mechanism and reflect the value of each option irrespective of the actual choice (FIGURE 35, top). The value is expressed as formal economic utility. It concerns specifically the value of a reward object (object value) and the value obtained by an action (action value). Although the decision process cannot change what is on offer, it maximizes the value by selecting the best object or action from the existing inputs (FIGURE 35, middle). This process requires sensory evidence for identification of each option as prerequisite for subsequent valuation of the options. Sensory evidence itself is not a decision variable in economic decisions, as it is imposed by the environment, and controlling it would not necessarily lead to the best reward. Furthermore, decision variables are closely related to the outputs of the decision process. The outputs controlled by the decision maker concern the reward value predicted from the decision (chosen value), the object that conveys that value (chosen object), and the selection of the valuable action (chosen action). Preparation, initiation, and execution of the action for obtaining that value are downstream processes (FIGURE 35, middle and bottom).

The value in the decision variables is rarely innate but acquired and updated using various forms of learning, in particular Pavlovian processes, operant habit learning, and operant goal-directed behavior, as described above. Through these three forms, objects and actions acquire formal economic utility, using model-free reinforcement learning as most basic mechanism, as stated in Equations 10B, 10E, 26A, and 26C. More complex learning and updating of utility may use various forms of model-based reinforcement learning.

Perceptual decision models, which are crucial for the sensory identification of reward objects (FIGURES 1, left, and 35, top), assume the gradual accumulation of evidence that needs to be extracted from noisy stimuli. Two main types of perceptual decision models vary in the way the inputs are compared with each other. Random walk, drift, or diffusion models have a single integrator with separate decision thresholds specific for each option (FIGURE 36A). Evidences are continuously compared between options through mutual inhibition at the input level (254, 302, 447, 454, 549, 565). The resulting difference signal grows and diverges over time towards one of the decision thresholds. The decision is executed for the option whose threshold is reached first. In contrast, race models have distinct accumulators for each option, each one with its own decision threshold. The evidences grow separately until the first option reaches a specific threshold, which then generates the corresponding behavioral choice (FIGURE 36B) (327, 614, 622). Race models are appropriate for choices involving several options that are tracked separately until a final comparison can be made. Intermediate forms between diffusion and race models differ in architecture and parameter setting (55). The valuation of each option (FIGURES 1, right, and 35, top) may occur as soon as the perceptual decisions have been made and each reward object has been identified. In other cases, the value may not be immediately apparent and requires some time to evolve into a stable representation. Only then can an economic decision variable arise. Thus a general formulation of economic decision mechanisms should include accumulation processes for sensory and value evidence.
The basic mechanism of the decision process is competition between independent options and underlies a large variety of decision models (55, 56, 441, 585, 623). It would mediate the on-going comparison and divergence of evidence in the diffusion models (FIGURE 36A), achieve the final comparison between options in the race models (FIGURE 36B), or constitute a decision mechanism on its own when evidence is immediate and accumulation negligible. The minimal model comprises two inputs that map onto two outputs while inhibiting each other’s influence (FIGURE 36C). Through forward excitation coupled with mutual lateral inhibition, the stronger neuronal signal becomes even more dominant by not only inhibiting more the weaker signal but also by being less inhibited by the weaker input. The mechanism amplifies the graded differences between the inputs, analogous to lateral inhibition of sensory systems, and, with various alterations, forms the basis for competition in a wide variety of diffusion and race models for perceptual and motor decisions (55, 56, 441, 585, 622, 623) and may also apply to economic decisions. Recurrent excitation would generate gradually increasing evidence accumulation in components of the network (622, 623). An additional threshold produces a winner-take-all (WTA) mechanism by removing the weaker signal and turning the graded difference into an all-or-none output signal that reflects only the value of the strongest option or consists of an all-or-none decision signal for the winning option. This comparative mechanism provides definitions for the major decision variables.

Neuronal decision signals should reflect the decision variables defined by the decision mechanisms. The art is to identify formal, biologically plausible decision models that employ the mechanisms defining the decision variables and are implemented with identifiable neuronal activities during various trial epochs (FIGURE 20, A AND D). The activities concern the initial accumulation of evidence, implemented by gradually increasing ramping activity in perceptual decisions (FIGURES 35, top, and 36, A AND B), and the subsequent valuation of the options. The central competitive process and its outputs (FIGURES 35, middle and bottom, and 36C) are reflected in distinct neuronal activities coding the input decision variables of object value (405) and action value (500), the abstract decision output (188), and the chosen value (405), chosen object (38), and chosen action (543). Updating of decision variables via prediction errors occurs globally via dopamine neurons and locally through specific groups of non-dopamine reward neurons (FIGURE 35, right). Neuronal decision signals are typically investigated in choice tasks, although basic processing aspects may be studied in imperative, forced-choice trials, which allow to assign neuronal coding to individual choice options and to predictions from a single stimulus. The following sections are devoted to these descriptions.

3. Evidence and valuation

The initial stage of the decision process concerns the rapid or gradual acquisition of evidence and the valuation of the
identified options (FIGURE 35, top). Acquisition of evidence takes time, in particular when the evidence itself is noisy. Neurons in many brain structures have the capability to bridge time gaps and accumulate information through persistent activity that may ramp up to future events. Ramping activity is one of the major neuronal mechanisms underlying decisions. This form of neuronal activity concerns sensory information in perceptual decisions, motor processes in action decisions, and economic utility with all its components in economic decisions (Equations 9–13, 16, and 21–25).

Ramping activity reflects a general propensity for persistent, gradually increasing activity in many cortical and subcortical neurons. It occurs without choice during several processes. During the expectation of sensory events, activity increases gradually toward predictable visual stimuli in premotor cortex (FIGURE 37A) (350), prefrontal cortex (65), parietal cortex (248), and striatum (6, 18, 215, 526, 528). During the instructed preparation of eye and arm movements, activity ramps up and differentiates between movements in primary motor cortex (467), premotor cortex (FIGURE 37B) (298, 491, 493, 636), frontal eye field (74, 506) [and its projection neurons to superior colliculus (147)], supplementary eye field (101, 506), supplementary motor area (491, 493), parietal cortex (109, 184), superior colliculus (379), and striatum (FIGURE 37C) (526, 528). Ramping activity to a threshold occurs also several seconds in advance of internally chosen, self-initiated movements in neurons of premotor cortex, supplementary motor area, parietal cortex, and striatum; these neurons are typically not activated with externally triggered movements (FIGURE 37D) (299, 311, 337, 380, 397, 491, 493, 526, 528). Similar human encephalographic ramping activity resembling the readiness potential is the basis for an accumulator decision model for movement initiation (515). During reward expectation, ramping activity occurs in orbitofrontal cortex (216, 544, 602), dorsal and ventral striatum (FIGURES 32C AND 37D) (18, 205, 215, 523), and amygdala (FIGURE 37E) (40). The anticipatory ramping activity seen with sensory events,

FIGURE 37. Neuronal ramping activity preceding stimuli, action, and reward. A: gradual activity increase during the expectation of an initial visual instruction stimulus in premotor cortex neuron. Open arrow points to movement triggering stimulus. [From Mauritz and Wise (350), with kind permission from Springer Science and Business Media.] B: gradual increase of left-right differential activity during movement preparation in premotor cortex neuron. Activity terminates just after the movement triggering stimulus and before the movement. Left (black) and right (gray) refer to target positions. [From Kurata and Wise (298).] C: ramping activity preceding self-initiated movements in striatum neuron. Activity begins gradually without external imperative stimuli and terminates with movement onset. [From Schultz and Romo (528), with kind permission from Springer Science and Business Media.] D: differential reward expectation activity in striatum neuron. Activity is lower in anticipation of raspberry juice (black) than blackcurrant juice (gray). [From Hassani et al. (205).] E: reward expectation activity in amygdala neuron reflecting instantaneous reward probability. Activity ramps up steeply before singular reward occurring at predictable time after stimulus onset (increasing reward expectation), but shows lesser increase before reward occurring pseudorandomly during stimulus (flat reward expectation with flat occurrence rate). [From Bermudez et al. (40).]
actions, and reward may constitute a basic neuronal mechanism for perceptual decisions, economic decisions, and internal movement generation.

During perceptual decisions, primary sensory systems code evidence about physical stimulus parameters irrespective of the subject’s perception and decision (120), whereas activity in secondary sensory cortex and premotor cortex reflects the animal’s subjective perception (120, 121). Ramping activity occurs over several seconds in neurons in parietal cortex, frontal cortex, and superior colliculus during the gradual accumulation of noisy evidence and reaches a criterion threshold at the moment of choice. Ramping activity is absent when sensory evidence is immediate and thus may not simply reflect neuronal processing (334). Ramping activity occurs during dot motion discrimination in prefrontal cortex, lateral intraparietal cortex (LIP), and superior colliculus (FIGURE 38A) (230, 274, 483, 535); length and brightness discrimination in superior colliculus (448, 450); and visual search and visual stimulus identification in frontal eye field neurons (FIGURE 38B) (49, 587, 588). Although typically tested with two choice options, ramping activity in LIP occurs also with four options (97). The slope of the ramp buildup to criterion threshold varies in LIP and superior colliculus with the speed and difficulty of the decision derived from the coherence of moving dots (FIGURE 38A) (483), differences in reaction time (97, 270, 448, 483), and differently rewarded options (494). The height of activity at choice threshold is the same irrespective of sensory coherence and motor response time in LIP (FIGURE 38A) (483). The statistical variance of neuronal activity in LIP increases gradually and declines earlier for one option compared with the other option immediately before the action (96). The probability of accurately predicting the animal’s behavioral choice from neuronal activity in LIP increases during the ramp (535), suggesting that intraparietal activity is suitable for affecting subsequent neurons involved in expressing the decision through action. Accordingly, ramping activity in prefrontal neurons changes its direction while monkeys change their decision (269). Choice accuracy depends on the starting point of the neuronal ramp but not its threshold in LIP (199). The ramping activity in perceptual choice tasks does not correlate with objective sensory stimulus parameters but reflects the subjective perception expressed by the animal’s choice, as shown with errors in direction judgement in LIP neurons (FIGURE 38C) (270, 535). Ramping activity may not always reflect motor processes, as it occurs in the frontal eye field even without action (587, 588) and in LIP correlates with motion strength evidence but not action (535).

During action decisions, even when straightforward sensory signals do not require lengthy accumulation of information, ramping activity in frontal eye fields builds up to a threshold for movement initiation before saccadic choices (198) and before correctly countermanded saccades (FIGURE 39A) (507). The rate of neuronal build-up determines saccadic reaction time (FIGURE 39B) (198, 508). Neuronal activity in cingulate motor area, frontal eye fields, and LIP ramps up before monkeys change movements to restore declining reward (FIGURE 39, C AND D) (101, 543). During delayed action instruction, neurons in premotor cortex

![FIGURE 38](http://physrev.physiology.org)
show initial nondifferential activity that becomes stronger, selective, and ramps up for the action once the target is revealed (FIGURE 39E) (98). As these activities occur in motor areas, reaching the threshold may reflect the completed decision process or the initiation of action.

With economic decisions between differently rewarded options, neurons in LIP show ramping activity with noisy reward options (494). This activity likely reflects the accumulation of sensory evidence before the reward valuation rather than constituting the valuation process itself. However, LIP neurons show ramping activity also with immediately evident, non-noisy information about the reward options, such as differently colored lights at specific spatial positions (FIGURE 40) (571). The neuronal activity scales with the economic decision variable of fractional income and predicts the chosen action. Neurons in anterior cingulate cortex show ramping build-up towards a single threshold when animals leave a foraging patch with decaying reward value (209). Steeper ramp slopes reflect shorter reaction times for leaving the patch, as the overt choice is initiated when the common threshold is reached. These ramping activities cannot reflect noisy evidence but may constitute components of the neuronal decision process in the transition from sensory processing to reward valuation. They might also reflect associated processes such as stimulus or reward expectation or movement preparation. Nevertheless, ramping activity is not a necessary feature for economic decision activity in frontal cortex neurons (404), although its absence in a studied neuronal pool would not contradict ramping activity in other, unrecorded neurons, as cortical neurons show vastly inhomogeneous task relationships. In a probabilistic choice task with unambiguous visual symbols, parietal cortex activity reflects the economic decision variable of log likelihood ratio of probabilistic rewards (643). A potential ramp is due to averaging of asynchronous, nonramping activity. Without outright decisions, responses of amygdala and dopamine neurons reflect the
transitions from initial sensory processing to reward valuation (Figure 9) (159, 422), which is considerably prolonged when the sensory evidence is noisy (389). Neurophysiological dopamine responses show a ramp with reward risk (Figure 31, C AND D) (161) but not with reward value. Striatal dopamine release ramps up before reward (237) which likely reflects presynaptic influences from reward expectation related cortical or amygdala inputs to striatum (40, 216, 544, 602) or from striatal neurons (18, 205, 215, 523) (Figures 32C and 37, D and E), but not a dopamine impulse ramp which does not occur with riskless rewards.

The morphology of ramping activity corresponds to the number of thresholds per neuron in the different decision models (56). The two-threshold drift-diffusion model assumes neuronal activity that ramps up differentially towards upper and lower thresholds for respective options (Figure 36A) (254, 302, 447, 565). Such activity is seen during perceptual decisions (Figure 38, A AND C) (97, 449, 483, 535) and economic decisions (Figure 40) (571). The observed gradual divergence of ramping activity in the same neuron may represent the output of an ongoing graded comparison with mutual inhibition in competing presynaptic neurons (Figure 36C). The single-threshold type of decision model conceptualizes a gradual divergence of ramping activity in the same neuron. The ramp either reaches the threshold for a specific option or fades back to baseline for the alternative option (Figures 36B). Such activity would also form the basis for race models in which activities from different pools of differentially active neurons race toward their individual thresholds (609, 614). Differential ramping activity is seen during the expectation of sensory events (Figure 37A) (6, 18, 65, 215, 248, 350, 526, 528), motor preparation (Figure 37B) (74, 98, 101, 109, 147, 184, 298, 299, 311, 337, 379, 380, 397, 467, 491, 493, 506, 526, 528, 636), reward expectation (Figure 37, D AND E) (18, 40, 205, 215, 216, 523, 544, 602), perceptual decisions (Figure 38, A AND B) (274, 483, 588), motor decisions (Figure 39A) (198, 507, 508, 587), and economic decisions (Figure 40) (571). The final step to the choice in these models may involve a graded or WTA competition between different neurons (Figure 36C).

Taken together, neuronal ramping activities constitute key mechanisms in the identification of choice options from sensory evidence and may partly underlie their valuation. They reflect both the evidence that the decision maker cannot influence and the decision variables controlled by the decision maker. Ramping activities resemble statistical decision processes for accumulating stochastic information about sensory evidence (185) and rewards and follow major assumptions of decision models (Figure 36, A AND B). The activities are widespread in the brain and reflect internal processes related to expectation of events, decisions between options, and preparation of action. As they occur also without choices, they may represent a general feature of neuronal physiology that is used for decision-making. Gradually increasing activity provides a focusing influence on downstream neurons through greater signal strength and higher signal-to-noise ratio, allowing better selection of behavioral actions (Figure 36C).

4. Object value

Object value denotes the value of a specific reward object. It derives intuitively from nutrient or other survival values of rewards that are sought by humans and animals for survival. In human societies, object value is fundamental for economic exchanges of goods in which a quantity of one good with a specific value is traded against another good with a specific value. Object value is the basis for the marginal rate of substitution in economics which is the amount of good A a consumer is willing to give up to obtain one unit of good B, utility remaining constant across the exchange (341). To choose optimally, we assess the utility of each object separately, compare between them, and then select the best one. The value updating after each experienced outcome involves reinforcement learning (Equations 10B, 10E, 26A, and 26C) with Pavlovian, habit, and goal-directed processes. Thus object value is defined as an input variable for competitive decision mechanisms (Figures 35, top, and 36C), in analogy to action value conceptualized in machine learning (575).

Object value has three crucial characteristics. First, it varies with the values of a reward object and, in principle, should be quantified as formal economic utility (Equations 9–12). Second, the value refers to one specific choice object. When choosing between several reward objects, such as a piece of meat and a loaf of bread, each object has its own and inde-

![Figure 40. Neuronal ramping activity during economic decisions. Ramping activity in monkey lateral intraparietal cortex neurons during oculomotor matching choice between more and less frequently rewarded options (average from 43 neurons). Activity reflects the "fractional income" (ratio of chosen reward value to summed value of all options). Blue, saccade into neuronal response field; green, opposite movement. Fractional income varies between 0 (dotted lines) and 1.0 (think lines). [From Sugrue et al. (571). Reprinted with permission from AAAS.]](image-url)
dependent value. “Objects” are solid or liquid rewards indicated by their appearance (color of fruit or fluid) and specific, intrinsically arbitrary stimuli associated with rewards. Different containers of the same kind of milk are examples of stimulus objects. Without being outright physical objects, events, situations and activities endowed with value have functions analogous to objects for economic decision mechanisms. The objects gain value through the animal’s actual experience with the reward or from inferences based on non-Bayesian or Bayesian models of the world. Third, object value is independent of the actual choice of the object. It indicates the value of a good irrespective of whether we will imminently receive or consume it, thus complying with the notion of reward information, rather than explicit reward prediction. Economic exchanges are based on value comparisons irrespective of immediate consumption. A pound of meat has an object value of $15 irrespective of whether I buy it now or not. Without this property, object values cannot serve as independent inputs to comparative decision mechanisms. Thus minimal tests for object value require two different reward objects, two values separately for each object, and independence from object choice.

Object value neurons process the value of one particular object irrespective of the animal’s action. There would be grape juice neurons tracking the value of grape juice irrespective of the animal’s choice, or banana neurons indicating how much the monkey will get if it were to climb the tree. The separate neurons tracking value for the specific objects constitute inputs for competitive decision models. In a binary decision, each input neuron, or pool of neurons, coding the value of its specific object competes with the input neurons coding the value of the other object. Hybrid subforms may show more pronounced object value coding if one particular object or action is chosen (combination with chosen object or chosen action coding defined below). To make meaningful comparisons, their signals need to vary on a common scale, both between the rewards (“common currency”) and between the neurons. Neither sensory differences between rewards nor different coding ranges or slopes between neurons should affect object value signals. The model can be easily extended to three and more options as long as each neuron codes only the value of a single object. Thus object value coding on a common scale appears to constitute a suitable neuronal decision signal for competitive decision mechanisms.

Neurons tracking the value of specific reward objects irrespective of reward choice or reception are found in orbitofrontal cortex. They code selectively the value of a specific fruit juice irrespective of the monkey’s choice and the required action [named “offer” value (405)] (FIGURE 41, A AND B) and thus do not predict the choice (404). These activities comply with the competitive formalism that requires independent coding of object value at the input (FIGURE 36C). Orbitofrontal object value coding remains selective with three rewards (406), conforming coding specificity for a particular object. Impaired object-reward associations during choices in orbitofrontal lesioned humans and monkeys underline an orbitofrontal role in object value coding (84, 108).

In contrast, most conventional reward responses reflect explicit reward prediction or reception and thus do not have the necessary property of coding object value irrespective of the actual choice, without which they could not serve as independent inputs to competitive decision mechanisms. They reflect the actual, free or forced, behavioral choice and probably code forms of chosen value and reward prediction (see below).

5. Action value

All choices lead to movements. Whether they are ocular saccades, licking, reaching, locomotion, or speech, they all consist of muscle contractions performed to get the best value for the effort. Even verbally expressed choices, and speaking itself, involve muscle contractions (of the larynx, mouth, and hands). Machine learning and reinforcement learning, which are concerned with artificial agents moving about the world, use action value to model the agent’s behavior for maximizing reward (575), like robots deciding which car part to weld first. Animals deciding between foraging patches are likely more interested in the value the action can bring them rather than considering each patch as an object. Each action leads to a reward value if it were chosen, thus the shorthand term of action value. Thus, for maximizing utility, an agent selects the action by which they achieve the highest estimated value. Pavlovian, habit, and goal-directed processes provide mechanisms for setting up and updating action value which, in principle, should be quantified as formal economic utility (Equations 9–25).

Actions are not only means for obtaining reward but may be themselves pleasurable and rewarding. The intrinsic reward may be added to the action value that usually refers only to the reward obtained from the action. The subjective nature of intrinsic action value emphasizes the need for quantitative behavioral assessments of often nonlinear and even nonmonotonic subjective value functions rather than inferring value from objective reward and action properties.

Like object value, action value is defined as input variable for competitive decision mechanisms (FIGURES 35, top, and 36C). Similar to object value, it requires three conditions. First, action value reflects reward value (amount, probability, EV, EU; Equations 9–13, 16, and 21–25). Second, value coding is specific for a particular action. The value is attached to an action and varies only for that action, such as an arm movement to the right or the left (spatial differential), an arm or an eye movement (effector differential), or going to work or vacation. Action value may arise from operant conditioning following the animal’s actual experi-
ence with the reward or from inferences based on models of the world, e.g., Bayesian. Third, action value is independent of the actual action choice and the value obtained by that choice. Without this property, action values cannot serve as independent inputs to competitive decision mechanisms. Thus minimal tests for action value require two different actions, two different values for each action, and independence from choice.

Action value neurons process the value for one particular action irrespective of the animal’s action. There would be left action value neurons tracking the value of a movement to the left irrespective of the animal’s actual movement on that trial, and there would be separate right action value neurons. Or different neurons would respectively track the value of arm and eye movements irrespective of the animal’s actual movement selection. Such separate action value tracking neurons are suitable inputs to competitive decision mechanisms that underlie utility maximization (FIGURE 36C). Subforms may show preferential action value coding for particular chosen actions or objects. Action value signals need to vary on a common scale, both between the actions and between the neurons, to constitute comparable neuronal decision signals. Sensory and motor parameters of the action should not affect action value signals.

Neurons coding action values according to these definitions are found in monkey and rat dorsal and ventral striatum (slowly and tonically firing neurons), globus pallidus, dorsolateral prefrontal cortex, and anterior cingulate of monkeys and rats. These neurons code selectively the value of the left or right arm or eye movement irrespective of the animal’s choice (FIGURE 41, C AND D) (245, 267, 275, 306, 500, 530). Action value neurons are more frequent in monkey striatum than dorsolateral prefrontal cortex (530). Action values in these studies are subjective and derived from behavioral models fitted to the choices (245, 267, 306, 500, 530) and from logistic regressions on the animal’s choice frequencies (275). However, the required common scale coding is unknown. The anterior cingulate cortex is involved in action-reward mapping during choices, as lesions...
there reduce preferences for more rewarded movements in humans and monkeys (84, 108).

The substantial action value coding in the striatum may reflect the pronounced motor function of this structure. Indeed, optogenetic stimulation in mice of dopamine D1 receptor expressing striatal neurons (direct pathway) induces within a few minutes behavioral preferences toward an operant touch or nose poke target associated with the stimulation, whereas stimulation of D2 receptor expressing striatal neurons (indirect pathway) induce behavioral dispreference (293). These data are compatible with the notion of striatal action value signals constituting inputs to a competitive decision mechanism located in the striatum or downstream structures.

In contrast, some reward signals reflect the explicitly predicted reward for specific actions while the actions are being planned or executed. These signals do not comply with action value coding, as they don’t show the required independence from free or forced choice and action and thus are unsuitable inputs to competitive decision mechanisms (see below under “chosen value”). They code the chosen value during choices and reward prediction in choices or imperative trials. If the signals code in addition specifics of the actions they may reflect goal-directed mechanisms (see above and FIGURE 20, E AND F).

6. Object value versus action value

The two decision variables derive from the intuition of objects having reward value that can be traded and require actions for obtaining them. The two variables may play different roles depending on the choice situation. Object value computations are likely to play a major role when values of objects change frequently and need to be tracked before choosing the required action. Once the object has been selected, its value can be directly mapped onto the action, without computing and comparing all action values. When I like to select a fruit at a new lunch place, I check the values of all available fruits (object value), and then I choose the highest valued fruit. Then I will know the action that gets me the chosen fruit (map the object to the action), without need to value each action and choose between their values. I can enjoy the feeling to soon get my beloved fruit, but that comes after the decision has been made and thus does not enter the competitive decision process. Indeed, monkey orbitofrontal neurons code the chosen value and map it onto action activity in dorsolateral prefrontal cortex to initiate the action (“good-to-action transformation,” Ref. 81).

In contrast, when object values are rather stable but actions change frequently, action value would be the more appropriate decision variable. With frequent action changes, effort cost may change and should to be tracked continuously and subtracted from income value (Equation 13). At the time of choice, the different object values are mapped onto action values while subtracting effort cost of the currently required actions. In realistic choice situations, the progression from object value to action value may not be so stereotyped, nor would object value have primacy over action value or vice versa. Rather, object values, action values, object-action mapping, and motor plans would evolve with partial overlap depending on the choice situation. Neuronal processing of these variables and mechanisms, and the competition mechanisms at each stage, may involve separate neuronal populations in different brain structures. Indeed, orbitofrontal and anterior cingulate lesions induce differential deficits in object-reward and action-reward associations (84, 108).

7. Abstract decision

Once the comparison process has resulted in selection of the best reward object, the output of the selection process may be specified by an abstract decision signal that concerns only the decision and as such may precede the specification of the choice. Abstract decision signals code the choice in a binary, categorical manner and irrespective of value, object, or action.

Abstract decision signals are seen during perceptual decisions. They become apparent when decoupling the decision temporally or spatially from the subsequent reporting action. The signals occur without, separately from, or stronger than object or action signals in prefrontal, premotor, secondary somatosensory, and anterior cingulate cortex with vibrotactile stimulus comparisons (Figure 42A) (212, 312, 334, 340, 489, 490) and visual stimulus detection (360, 361). They are also seen as buildup activity in supplementary eye fields and superior colliculus during visual search (587) and random dot motion discrimination (230).

Abstract decision signals occur also in specific neurons in the amygdala that show graded reward value coding early during the trial and transition to later binary abstract decision coding (Figure 42B) (188). The decision signal from 1–50 pooled neurons predicts the upcoming choice correctly in 75–90% of trials, respectively, and remained predictive even with identical reward between the two options. These neurons are not activated during imperative forced-choice trials, nor do they code visual stimuli indicating the options or the subsequently chosen oculomotor actions. Abstract decision signals during economic choices in the amygdala extends its recently recognized reward function (422) to economic decision-making and challenge the view of a structure primarily known for fear responses.

8. Chosen value

Chosen value reflects the value of the specific object or action that the decision maker obtains from the choice. Its
function is defined by competitive decision mechanisms (FIGURES 35, middle, and 36C). Whereas object value and action value are attached to an object or action irrespective of the choice, chosen value reflects the explicitly predicted or received reward value following the choice of that object or action. Thus chosen value reflects the output of the decision mechanism. For this important difference, chosen value should not be called object value or action value, even when it is attributed to a specific object or action. Chosen value is a decision variable, as it is controlled by the decision maker, who usually tries to maximize it (with the exceptions mentioned above). Chosen value, in principle, should be quantified as formal economic utility (Equations 9–25). Reinforcement learning employing Pavlovian, habit, and goal-directed processes provide mechanisms for setting up and updating chosen value.

Chosen value comes in two forms. Chosen object value reflects the value of a chosen reward object, like apple or orange, or the value of a specific stimulus predicting a particular reward. In contrast, chosen action value reflects the value of a chosen action. This distinction can be exemplified with particular reward magnitudes and probabilities. Rewards A and B are 1 ml and 2 ml of juice, respectively. Reward A occurs at the left in 20% of trials and at the right in 80%, whereas reward B occurs at the left in 60% and the right in 40% of trials. Then chosen object values are 1 ml and 2 ml for rewards A and B, respectively, whereas chosen action values are 1.4 ml for left (1 ml × 0.2 + 2 ml × 0.6) and 1.6 ml for right action (1 ml × 0.8 + 2 ml × 0.4). Most experiments do not specifically address this distinction by not using such discriminating values.

Neuronal chosen value signals reflect the value of an object or action irrespective of nonvalue attributes. These signals do not code the kind of reward (sensory aspects), sensory properties of predictive stimuli, spatial reward occurrence, stimulus positions, and required actions. In free choices between objects, specific neurons in orbitofrontal cortex code the chosen value of two or three different reward juices (FIGURE 23, A AND B) (405, 406). These activities reflect the estimated or experienced subjective economic value of the chosen juice, as assessed psychophysically by behavioral choices, and thus predict the choice (404). The signals scale in the same way across different rewards, thus coding value in common currency. Dopamine neurons code chosen value irrespective of reward probability associated with the chosen stimulus (377). Stronger dopamine activations following conditioned stimuli predict the choice of the higher valued stimulus. Chosen value signals in the striatum show temporal discounting in choices between identical reward liquids delivered after different delays (79). Chosen value signals in anterior cingulate cortex code reward magnitude, probability, and effort, often in the same neurons (261). Chosen value signals in the supplementary eye field reflect the subjective value of objects, as estimated from certainty equivalents of behavioral choices between certain and risky outcomes (551). In imperative forced-choice trials, neurons in striatum and substantia nigra pars reticulata distinguish between stimuli predicting different reward magnitudes irrespective of sensory stimulus properties (107, 654).

In free choices between actions, some striatal neurons code chosen value without coding the action itself. The value signals vary with the reward probability associated with the chosen action and predict choice of the better action with higher activity during several task epochs (275, 306). Some striatal reward magnitude coding neurons do not distinguish between actions in imperative trials but reflect the reward value predicted from the choice (107). Chosen value signals in the striatum often occur after the overt behavioral choice, whereas action value signals typically appear before the choice (275, 306), suggesting that these chosen value signals represent the result of the decision process.

Unchosen value refers to the value of an unexperienced outcome. Whereas chosen value reflects the actual reward and is appropriate for experience-based decisions and reinforcement learning, unchosen value reflects some knowl-
edge about the world and might be more appropriate for model-based behavior. Animals benefit from the unchosen value information to improve reward gains (1, 208). Some neurons in dorsal and ventral striatum code unchosen value during intertemporal choices in monkeys (79) and during locomotor choices in rats (275). These neurons are distinct from the more frequent chosen value coding neurons in these structures. Some neurons code both chosen value and unchosen value with similar response slopes in anterior cingulate (208) and in orbitofrontal and dorsolateral prefrontal cortex (1). Thus some chosen value neurons code value irrespective of actual outcome.

Even before choosing, subjects may inspect the different options and process the anticipated value of each option. Humans show such inspectional eye movements whose frequencies and durations comply with evidence accumulation in diffusion decision models (291). It would be interesting to see whether neuronal value processing may initially reflect the value of the object or action being considered and subsequently shift to coding the chosen value. During action decisions, neurons in premotor cortex show initial activity for both actions but become selective for the final action after an imperative instruction (98, 99).

Hybrid forms of chosen value include chosen objects or actions. Corresponding neuronal signals reflect the value of the chosen reward together with sensory or motor information about the chosen object or action. If these activities occur well before the overt behavioral choice and increase towards a decision threshold, they may reflect the evidence accumulating in favor of the chosen option and thus play a role in the decision process. If the activities occur after the decision, they would constitute the outputs of the decision process and not drive the decision itself. They are occasionally referred to as object value or action value coding. However, according to the definition by competitive decision mechanisms (FIGURE 36C), they constitute chosen value coding, as they reflect the reward value derived from the chosen objects or actions. Neuronal activity combining expected reward value conjointly with object information is only reported from imperative trials in frontal and temporal lobe neurons, as mentioned above (231, 319, 371, 394, 411), but cannot be labeled as “chosen” value due to the nonchoice task nature. Neurons coding chosen value conjointly with action are found with free choices in dorsal and ventral striatum (275, 306, 418, 479); globus pallidus (418); premotor, prefrontal, and anterior cingulate cortex (210, 532, 550, 620); and parietal cortex (433). The actions are eye or hand movements towards different spatial targets in monkeys (210, 306, 418, 433, 494, 550, 571, 620, 643) and nose pokes and whole left versus right body movements in rats (275, 479). In striatum, hybrid chosen value and action coding neurons are distinct from neurons coding only chosen value (306) or action value (306, 572). Conjoint coding exists also between unchosen value and action in orbitofrontal and dorsolateral prefrontal cortex (1). Comparable conjoint coding of value and action occurs also in nonchoice tasks (107, 135, 178, 205, 221, 243, 260, 282, 308, 313, 348, 354, 381, 427, 476, 502, 542, 606, 627), without qualifying for chosen value coding.

9. Relative value

The competitive decision mechanism for maximizing utility (FIGURE 36C) does not depend on actual values but involves comparison of the options relative to each other and is thus based on the difference or ratio of their values. Thus relative value is an important decision variable which, in principle, should be quantified as utility. Random walk-diffusion models capture the importance of relative value by assuming an early comparison with subsequent evidence growing toward specific thresholds. The mechanism for coding relative value may involve mutual inhibition between inputs that generates a graded difference between the inputs, amplified by the inhibition exerted by stronger inputs onto weaker inputs.

For binary decisions between options 1 and 2, the relative value can be stated using the final utility defined in Equation 25

\[
\text{REU}_{\text{sum}} = \text{EU}_{\text{sum}1} - \text{EU}_{\text{sum}2} \quad (27)
\]

Instead of EU_{sum}, any of the constituent utility functions defined by Equations 9–13, 16, and 21–24 can be used for Equation 27. Most simply, \( \text{REU}_{\text{sum}} \) may be derived from objective magnitudes \( m_1 \) and \( m_2 \)

\[
R_m = m_1 - m_2 \quad (27A)
\]

to obtain relative utility

\[
R_u = u_1 - u_2 \quad (27B)
\]

By applying a logarithmic utility function \( u = \log (m) \), we obtain

\[
R_u = \log(m_1) - \log(m_2) \quad (27C)
\]

which is equivalent to

\[
R_u = \log(m_1/m_2) \quad (27D)
\]

As Equations 8 and 9 show, reward value derives not only from reward magnitudes but also from reward probabilities. We replace \( m_1 \) and \( m_2 \) by probabilities \( p_1 \) and \( p_2 \)

\[
R_u = \log(p_1/p_2) \quad (27E)
\]

\( R_u \) is a log odds ratio, also called log likelihood ratio (643). General simplifying formulations would replace objective \( m \) and \( p \) by their summed product of EV.

Another good measure of relative value is the ratio of objective values (magnitudes or probabilities).
The simple value ratio is often extended to include the summed value of all options in the denominator, called expected gain ratio (Equation 27F) or fractional value (Equation 27G):

$$R_S = \frac{v_1}{\sum v_i} \quad \text{for } i = 1, n; \quad n = \text{number of options}$$

Reference value defined by Equations 27-27G applies to all forms of economic value, including object value, action value, and chosen value, and to all value functions defined by Equations 8–13, 16, and 21–24. In monkeys, binary behavioral choices follow relative subjective value defined by the log odds ratio (Equation 27E) (643), reward magnitude ratio (Equation 27F) (306), or the ratio of objective chosen reward magnitude or probability over all options (Equation 27G) (433, 571). These measures are good predictors of choice preferences.

The difference value of Equation 27 is analogous to reference-dependent utility (Equation 22), as both terms indicate relative value. Thus difference value and value adaptation serve the same function. They allow comparisons between existing options rather than requiring separate computations of the values of all options. Value ratios of the chosen option over the alternative (Equation 27F) or all options (Equation 27G) add adaptation to the spread of the reward probability distribution and correspond to divisive normalization in adaptive coding (332).

Some neurons in monkey and rat dorsal and ventral striatum, globus pallidus, and anterior cingulate cortex code difference subjective action value during free choices with eye or body movements, conforming to Equation 27A (79, 245, 306, 531, 572). Subjective value is derived from reinforcement models fitted to the choices or logistic regressions on choices. During temporal discounting, neurons code differences in subjective action value in dorsal striatum, and to some extent in ventral striatum (79). The discounted, subjective value is assessed from logistic regressions on inter-temporal choices. Due to the relative nature of difference coding, neuronal activity varies positively with the value of one action or negatively with the value of the alternative action. True to the definition of action value, these relative value signals do not depend on actual choice. Neuronal difference signals exist also for chosen value in dorsal and ventral striatum, which take both chosen and unchosen values into account (79, 306). These activities increase with the value of one option and decrease with the alternative’s value and, different from action value, depend on the actual choice. Neuronal activity in parietal association cortex reflects the reward ratio of the chosen option over all possible options (Equation 27G) (137, 433) and reaches a threshold at the time of choice (Equation 40) (494, 571). Similar parietal activity follows the increasing, logarithmic reward probability ratio between two options (Equation 27E) in the statistical sequential ratio test (Figure 43A) (643). Finally, difference signals for hybrid chosen value and oculomotor action exist in dorsal and ventral striatum (79, 306). Taken together, there are groups of distinct reward neurons coding basically every form of value decision variable in a relative way.

10. Chosen object and action

Although economic decision variables center around value, once the decision has been specified internally, the process may transfer to the overt behavioral choice without any longer involving value. Some neuronal signals in principal reward structures code the chosen object or the chosen action (frontal cortex, parietal cortex, and striatum) (Figure 35, middle).

Chosen object signals code the specific reward selected during the decision process or the specific visual stimulus pre-
dicting a particular reward. They serve to identify the chosen reward object while it is being acquired by an action. The signal does not code value and is not a simple visual or somatosensory response, as it may precede the appearance of the object. With economic decisions, chosen object signals in orbitofrontal cortex reflect different juices (405). With perceptual decisions, LIP neurons code the target color chosen for reporting the earlier, temporally distinct random dot motion direction (FIGURE 44A, orange) (38). The signal is distinct from abstract decision and chosen action coded by other parietal neurons.

Chosen action signals code the specific action for acquiring the reward and may serve to prepare, initiate, and execute the necessary motor processes. These signals reflect the future action rather than reward value. They occur in cingulate motor area, anterior cingulate cortex, frontal and supplementary eye fields, and parietal cortex when monkeys change their actions after the reward for the current action declined (543), even before contacting a switch-over key (209) or before a chosen target occurs (FIGURE 39D) (101).

The signals are absent or less pronounced in imperative forced-choice trials (101, 543). Neurons in LIP code the chosen saccadic action during probabilistic reward choice (433, 571), intertemporal choice (477), and dot motion discrimination (494), both during free choices and imperative forced-choice trials. In monkey and rat striatum, different neurons code chosen action from action value or chosen value (571). Thus chosen action signals occur in reward neurons close to behavioral output.

With perceptual decisions, parietal cortex neurons code the saccadic action for reporting random dot motion direction, distinct from abstract decision coding (FIGURE 44A, blue) (38). Neurons in superior colliculus show chosen action signals superimposed on abstract decision signals (230).

With motor decisions, parietal cortex neurons show differential activity preceding freely chosen eye or arm movements (FIGURE 44B) (109, 509). During instructed action selection, separate neurons in premotor cortex initially code all options, but their activity differentiates and ramps up specifically for the arm movement imposed by a cue (FIGURE 39E) (98, 99). Although not reflecting free choice, the activity shows a nice transition from separate option coding to specific option coding as a general template for the progression of neuronal coding of economic decision variables.

11. Updating economic decision variables

Whereas perceptual decisions are based on sensory evidence, economic decisions require additional information about the value obtained by the choice. For meaningful decisions, this information needs to be predictive of the outcome. Choices without predictions amount to guessing. The value prediction does not derive from sensory organs but is learned and updated by error driven reinforcement learning. Its crucial error term is computed as the difference between the experienced reward and its prediction (Equations 1, 2, 4, 5, 8A–D, and 26–26C).

The prediction entering the error computation derives from different sources (575). In model free reinforcement learn-

![FIGURE 44. Chosen object and action signals in monkey lateral intraparietal cortex. A: coding of chosen object (orange color) or chosen action (blue) during perceptual decisions in population (averages from 23 neurons). In the conditional motor task, the animal saccades to a red or green dot (chosen object) depending on the direction of random dot motion. The dots are shown at different spatial positions and thus require different saccades (chosen action). Ordinate is % of significantly responding neurons for chosen object and chosen action (and dot motion direction, magenta). [From Bennur and Gold (38).] B: chosen action coding in single neuron. After a common onset, activity differentiates between effectors ~400 ms before action (saccade vs. reach). [From Cui and Andersen (109), with permission from Elsevier.]
ing (460), the prediction derives only from direct experience with the values of the chosen objects or chosen actions in the past several trials. Other decision variables providing predictions, such as object value and action value, are only partly suitable for computing prediction errors because they are independent of the object or action being chosen and thus independent of the actually experienced reward. Their predictions are only useful for computing prediction errors when the objects or actions are chosen and their outcomes experienced, whereas unchosen rewards cannot enter the error computation. Thus the prediction in standard, model-free reinforcement learning derives from chosen value, which is a decision variable. In model-based reinforcement learning, the prediction incorporates information from models of the world and does not only reflect the recent experience with the outcome. The value prediction can be accurate even for objects and actions that have not recently been chosen and experienced. For example, given a model that assumes a constant sum of left and right action values, the left action value prediction can be accurate after only experiencing a right action and its outcome. Model-based reinforcement learning incorporates information beyond the experienced chosen value into the prediction and thus affects the error computation, whereas the reward is experienced as a matter of fact irrespective of model or not.

The type of prediction entering the error computation affects the form of the reward prediction error and thus the efficacy of value updating in specific learning scenarios (575). Three models are noteworthy in this respect. First, the Sarsa model applies to a range of laboratory tests in which the agent can experience all options during learning. Its prediction error incorporates the prediction derived from chosen value, which reflects the choice, but prediction from object value and action value is only valid when it derives from the actual choice. Thus chosen value neurons in striatum, orbitofrontal cortex, supplementary eye field, and anterior cingulate cortex (79, 208, 245, 261, 275, 306, 377, 405, 551) may serve as prediction inputs for the Sarsa model, and dopamine neurons code prediction errors in chosen value according to the Sarsa 2 model (377). Second, Q learning directly approximates the optimal reward value within a given choice set. Its prediction error incorporates the prediction of the maximal attainable reward value, without taking the choice into account. Neurons in monkey anterior cingulate cortex code the prediction for the maximal probabilistic reward value that the animal usually but not always chooses (called task value) (12). Outright prediction error signals coding maximal value have not been described. Third, actor-critic models incorporate the sum of reward predictions from all options. Neurons in striatum, globus pallidus, and anterior cingulate cortex code the predicted sum (or mean with steeper slope) of reward values (79, 208, 245, 531). Taken together, the Sarsa reinforcement model has a correlate in the prediction error signals of dopamine neurons and other reward neurons that reflect predictions derived from chosen value. Future experiments might search for prediction error signals of the other reinforcement models in a wider range of learning situations.

Error signals suitable for updating decision variables are found in several reward structures. The bidirectional dopamine reward prediction error response (517) occurs with visual, auditory, and somatosensory stimuli (221, 365, 492) irrespective of action (521, 597, 618); integrates reward delay, reward risk, and different liquid and food rewards into a common subjective value signal (285, 301); incorporates model-based predictions (68, 382, 598); and reflects chosen value (377). Subgroups of non-dopamine neurons show bidirectional prediction error signals in lateral habenula, striatum, globus pallidus, amygdala, anterior cingulate cortex, and supplementary eye field (13, 36, 134, 227, 261, 275, 344, 531, 551). Some error signals in striatum code additional sensory and motor parameters (556). Neurons in cingulate cortex and supplementary eye field code bidirectional prediction errors from object value and action value to the extent of the object or action being chosen and leading to actual reward experience (chosen object value and chosen action value) (531, 550). Possible prediction error signals are also the separate, unidirectional positive and negative prediction error activations in anterior cingulate and supplementary eye field (246, 261, 551).

The updating function of the dopamine prediction error signal can be considered within the generic competitive model of decision-making (FIGURE 45). Irrespective of the differences between formal decision models (55, 56, 441, 585, 623), the dopamine response would primarily update the inputs to the comparison process, namely, object value and action value that are coded during various trial epochs (FIGURE 20, A AND D). The experienced and the predicted reward for the dopamine error computation likely derive from chosen value. The dopamine signal is rather similar among individual neurons and has heterogeneous effects on heterogeneous postsynaptic neurons. However, object value and action value are processed irrespective of the decision. They do not provide the necessary postsynaptic specificity to make the dopamine influence selective, which would indiscriminately update all object values and action values, including those not even experienced. However, as primary and higher order rewards occur only after an event, neuronal updating involves stimulus traces that maintain a label of the engaged synapses beyond the event and make them eligible for modification (574). Thus each object value and action value signal may leave a distinct eligibility trace that decays rapidly unless being stabilized by neuronal inputs (FIGURE 14, B AND C). Neuronal activity from the specific chosen object/chosen action signal may provide selective inputs to object value/action value neurons for the same object/action for stabilizing their eligibility traces (FIGURE 45, right), whereas the lack of activity from the unchosen option would not prevent the decay (left). Thus the
maintained eligibility trace reflects the output of the decision process and provides the necessary specific label for dopamine influences to act upon. As a result, the indiscriminate dopamine signal would selectively update the object value and/or action value of the chosen option, whereas object value and/or action value of the unchosen option remain unaltered and need to wait for their next selection before being updated (except in model-based reinforcement learning, see above). Thus the influences of chosen object and chosen action on the eligibility trace provide the necessary selectivity for dopamine action. As an example (FIGURE 45), choice of a movement to the right and experience of the corresponding reward would lead to updating of right action value by a prediction error signal. If the right choice resulted in a higher than predicted value and thus a positive prediction error, right action value would be increased; however, if the value is lower than predicted and elicits a negative prediction error, right action value would be decreased. If the right action is not chosen, its value cannot be experienced and thus will not be updated (or might decay depending on the algorithm implemented). Taken together, within a simple comparator model of decision-making, the global influence of the dopamine error signal derived from chosen value may serve to update input decision variables for specific options via eligibility traces selectively stabilized by activity from the chosen option.

Neuronal correlates for this hypothesized dopamine influence consist of the global projection of this reward signal to most or all neurons in striatum and frontal cortex and to many neurons in amygdala, thalamus, and other structures (FIGURE 17A). The released dopamine influences the plasticity between conjointly active presynaptic and postsynaptic neurons in a three factor Hebbian learning scheme (FIGURE 17, B AND C) involving striatum, frontal cortex, and other structures in which appropriate dopamine dependent plasticity is found (82, 194, 265, 294, 400, 401, 424, 461, 495, 540, 583, 642, 648). In this way, the dopamine signal could affect synaptic transmission onto striatal and cortical neurons coding object value and action value. These effects may be differential within striatal regions, as action value neurons are more abundant in dorsal compared with ventral striatum (245), whereas ventral striatum neurons are more involved in reward monitoring unrelated to action coding (245, 523). The dopamine plasticity effects may also be differential for striatal neuron types, as optogenetic stimulation of neurons expressing D1 or D2 dopamine receptors induces learning of behavioral preferences and dispreferences, respectively, possibly by updating action values (FIGURE 19F) (293). These neuronal data support the scheme of dopamine updating of decision variables (FIGURE 45).
Updating with non-dopamine error signals may be less global and more selective. These neurons likely exert focused influences on specific postsynaptic neurons through selective anatomical projections. Their error signals are more specific for the type of information they provide and for the type of postsynaptic neuron they influence. Thus they are likely to be more specific for the objects and actions and the type of neuronal value signals they update.

12. Immediate dopamine influences

In addition to affecting synaptic plasticity and learning, phasic dopamine signals exert immediate influences on postsynaptic processing and behavioral performance and choices. The anatomical basis may be the triad arrangement of dopamine synapses on glutamatergic inputs to dendritic spines of striatal and cortical neurons (FIGURE 17B) (171, 186) and convergence in striatum and globus pallidus (416, 430). The long debated focusing effect assumes that dopamine signals suppress the effects of weaker inputs to the striatum while allowing stronger ones to pass through to the basal ganglia outputs (70, 368, 594). Thus phasic dopamine signals may affect the function of existing anatomical convergence and modulate the use of neuronal connections. At the cellular level, the phasic dopamine signal has excitatory action on D1 receptor expressing striatal neurons of the direct pathway by prolonging transitions to membrane up states (depolarization) (213). In contrast, dopamine action on D2 receptor expressing striatal neurons of the indirect pathway reduces membrane up states and prolongs membrane down states (hyperpolarization) (214). In a hypothetical model of such action, the dopamine enhancement of cortical input efficacy may induce reward related activity in striatal neurons (FIGURE 46A). This influence may affect reward processing in all forms of striatal task related activity (FIGURE 20, A AND D). Impairments of dopamine transmission reveal the immediate dopamine influences. Striatal dopamine depletion reduces learned neuronal responses in striatum (13), alterations of D1 receptor stimulation affect mnemonic behavior and neuronal activity in frontal cortex (503, 616), and reduction of dopamine bursting activity via NMDA receptor knockout prolongs reaction time (650). These deficits relate well to the immediate behavioral effects of phasic dopamine stimulation. Optogenetic activation of mouse dopamine neurons, in addition to providing reinforcement, elicits immediate behavioral actions, including contralateral rotation and locomotion (271). Pairing electrical stimulation of the medial forebrain bundle containing dopamine axons with lateralized stimulation of somatosensory cortex induces forward locomotion in rats and directs the animals in sophisticated and even unnatural spatial navigational tasks (FIGURE 46B) (580). The forebrain bundle stimulation likely serves as immediate reward, whereas the cortical stimulation provides the necessary spatial information. Corresponding to the action-inducing effects of dopamine stimulation, indirect inhibition of VTA dopamine

FIGURE 46. Immediate dopamine influences. A: hypothetical mechanism of influence of dopamine reward signal on striatal reward response. Bottom: a rewarded, but not unrewarded, stimulus differentially activates a dopamine neuron. [From Tobler et al. (598).] Middle: possible dopamine influence on striatal synaptic potentials: dopamine D1 receptor agonist prolongs the depolarization of a striatal neuron. [From Hernández-Lopez et al. (213).] Top: as a result of the dopamine influence on striatal depolarization, the striatal neuron responds stronger to a rewarded compared with an unrewarded stimulus. [From Hollerman et al. (222).] B: influence of dopamine reward signal on behavioral navigation. Electrical stimulation to the left and right somatosensory cortex of rats provides cues for turning; electrical stimulation of the medial forebrain bundle containing dopamine axons induces forward locomotion. The combined stimulation is able to guide a rat through a three-dimensional obstacle course, including an unnatural open field ramp. Colored dots indicate stimulation of forebrain bundle and somatosensory cortex. [From Talwar et al. (580). Reprinted with permission from Nature Publishing Group.] C: influence of dopamine reward signal on behavioral choices. Unilateral optogenetic stimulation of mouse striatal neurons expressing dopamine D1 or D2 receptors induces immediate differential choice biases (left bias for right D1 stimulation, top; and right bias for right D2 stimulation, bottom). [From Tai et al. (576). Reprinted with permission from Nature Publishing Group.]
neurons via optogenetic activation of local, presynaptic GABA neurons reduces reward consumption (613). At the postsynaptic level, optogenetic stimulation of dopamine D1 receptor expressing striatal neurons increases behavioral choices toward contralateral nose poke targets, whereas stimulation of D2 receptor expressing striatal neurons increases ipsilateral choices (or contralateral dispreferences) (FIGURE 46C) (576), suggesting an immediate, differential effect on reward value for the two actions.

Whereas these effects are straightforward, stimulating a couple synapses upstream of dopamine neurons induces surprisingly complex effects. Optogenetic activation of telentorial inputs to dopamine neurons projecting to nucleus accumbens elicits place preference, whereas stimulation of D2 receptor expressing striatal neurons increases ipsilateral choices (or contralateral dispreferences) (FIGURE 46C) (576), suggesting an immediate, differential effect on reward value for the two actions.

The immediate behavioral dopamine effects may be captured by a simplifying account that incorporates the essentials of immediate dopamine influences on behavioral choices into a generic competitive decision model (FIGURE 47). To have an immediate effect, the dopamine signal

![Diagram](image-url)
should reflect the prediction error at the time of the decision preceding the final reward. This error is the TD reward prediction error between the current prediction reflecting the chosen value minus the preceding reward prediction. The signal may influence neuronal activity for two kinds of decision variables. A dopamine-focusing effect (70, 368, 594) on predecisional object value or action value signals would enhance the value difference between the two options in a graded manner. The indiscriminate dopamine signal would affect all active synapses on neurons coding by definition object value and action value irrespective of the decision. Such dopamine influences on action value coding were assumed to explain the differential effects of optogenetic stimulation of D1 and D2 containing striatal neurons on behavioral choices (576). A more divisive dopamine influence would be achieved by affecting neurons that code the postdecisional variables of chosen object or chosen action before the overt behavioral choice (FIGURE 47, right). Activity in these neurons is differential due to selection by the WTA decision mechanism. The nonexisting activity from the unchosen option would not be enhanced by the dopamine signal (left). Thus the WTA mechanism would provide selectivity for the indiscriminate dopamine influence. Taken together, the basic comparator model of decision-making provides an architecture possibly underlying the immediate influence of the phasic dopamine signal on decision processes. This hypothesis is intentionally crude and provides only a general account, without addressing details of implementation by different types of dopamine innervated neurons in striatum, frontal cortex, and amygdala.

13. Formal neuronal decision models

The remarkable correspondence of neuronal signals to the key features of decision models encourages the development of biologically inspired neuronal models of decision-making. In many of such models, the economic decision variables of individual choice options are processed in separate pools of spiking neurons. The ramping activity is generated from transient inputs by recurrent synaptic excitation involving AMPA or NMDA receptors to induce self-sustained attractor states and integrate sensory or value information. The competition between choice options is mediated by GABA neurons exerting mutual, lateral, feedback, or feedback inhibition between the neuronal pools for the choice options. The WTA mechanism combines the inhibition with a thresholding mechanism that leads to acceleration of ramping activity or to all-or-nothing selection between signals. The winning suprathreshold activity then propagates to pools of output neurons that mediate the behavioral action.

For neuronal decision models involving the cerebral cortex, key features are the discovery of explicit neuronal decision signals and the local synaptic connectivity. A generic model for perceptual discrimination of random dot motion direction in LIP neurons uses a difference signal derived from pools of specific direction-sensitive input neurons from middle temporal (MT) cortex (535). For example, larger activity in rightward compared with leftward motion coding MT neurons results in an LIP difference signal for rightward motion. The activity begins with appearance of the moving dots and accumulates over time with the continuing incoming noisy evidence about dot motion direction until it reaches one of the two decision thresholds. This model conforms to a diffusion model with continuous comparison between options and activity diverging towards one of several thresholds (55) and fits well the experimental data (535). Through its generality, such a model would also fit economic decision processes during which LIP neurons code the log likelihood ratio decision variable in a sequential probability ratio test (643), which constitutes an efficient decision mechanism (619).

A cortex model of two-stage somatosensory frequency discrimination uses two mutually inhibitory pools of 250 noisy spiking neurons that represent the two choice options (334). The model neurons show sustained activity that scales with the frequency of the first, transient vibratory stimulus until the comparison stimulus occurs. The subsequent activity conforms to an abstract decision signal that does not vary with individual stimulus frequencies. The sustained activity and its transition from mnemonic coding to decision signal resemble closely the activity of prefrontal and premotor cortex neurons of monkeys in this task (65, 212, 488, 489). The model achieves categorical decisions by discriminating between vibratory stimuli irrespective of their actual frequencies and thus supports the function of the experimentally observed neuronal signals in perceptual decisions.

More elaborate cortical neuronal decision models employ architectures comprising frontal and parietal cortex, basal ganglia, and superior colliculus (FIGURE 48A) (323, 622, 623). Transient inputs induce persistent activity with slow temporal integration via recurrent cortical NMDA and AMPA mediated excitation. Mutual lateral inhibition between pools of 240 irregularly spiking neurons mediates the competition between the stochastic choice options. By excitatory attractor dynamics and WTA competition throughout decision formation, the ramping activities in the neuronal pools representing the choice options become increasingly more separated in their race toward specific decision thresholds. The ramping activities reflect evidence accumulation or decision variables such as event likelihood ratio for perceptual decisions or object value, action value, and derivatives for economic choices. The cortical activity exceeding the competition threshold is detected by the superior colliculus via direct or double inhibitory influences through caudate nucleus and pars reticulata of substantia nigra and then propagated to the oculomotor system. The modeled neuronal activities replicate well the experimentally observed
ramping in frontal and parietal neurons described above. Dopamine-mediated three-factor Hebbian reinforcement learning may set the decision threshold in caudate nucleus or cortex (554). Thus cortical WTA determines the ramping, superior colliculus detects threshold crossing, and dopamine neurons set the threshold.

For neuronal decision models involving the basal ganglia, key features are the well-organized corticostriatal projections, internal connections, and parallel loops with cortex (7, 141, 164, 196, 197, 252); the synaptic convergence through the component structures (FIGURE 48B) (416, 430); the separate striatal outputs with sequential inhibitory neurons (93, 127); and the dopamine-mediated plasticity of corticostriatal connections (424, 517, 540). The basal ganglia may be involved in two distinguishable aspects of decision-making, namely, the direct selection of options and the updating of option values through reinforcement learning (57).

Option selection in the basal ganglia employs the dopamine focusing effect (70, 368, 594) and anatomical convergence in striatum and globus pallidus (416, 430) (FIGURE 47). The models employ physiological activities and known input-output relationships of basal ganglia nuclei to demonstrate the feasibility of movement selection (72). By implementing the efficient sequential probability ratio test, the scaled race model shows that the subthalamic nucleus modulates the decision process proportionally to the conflict between alternatives (57). It outperforms unscaled race models. Particularly good fits are obtained with a diffusion model of movement selection (449). The basal ganglia decision models are supported by experimental data not used for their construction and also elucidate economic decision mechanisms, including the differential effects of optogenetic stimulation of dopamine D1 and D2 receptor expressing striatal neurons (FIGURE 46C) (576). The opposing functions of the direct and indirect striatal output pathways are modeled by a race between movement and countermanding signals that reflect the relative timing of the two signals as they race towards a threshold (511). The signals fit experimentally observed activities in rat substantia nigra neurons. These studies provide key support for important basal ganglia contributions to motor selection and economic decisions.
Updating of object value, action value, and their derivatives by the basal ganglia employs dopamine prediction error signals (517) [FIGURE 45]. Data analyses routinely employ the modeled subjective values of choice options as regressors for neuronal value coding and action value updating in the striatum (245, 267, 275, 306, 500, 530). Reinforcement models distinguishing between direct and indirect striatal pathways are based on in vitro work (424, 540) and assume that dopamine prediction error signals have differential plasticity effects on D1 versus D2 receptor expressing striatal neurons (72) and induce the learning of respective choice preferences and dispreferences (169). These results from formal models closely reflecting physiological data confirm the capacity of dopamine reinforcement signals for learning and updating of economic decision variables.

E. Reward and Free Will

Current notions about free will range from extreme determinism, in which every event of every individual at every moment is predetermined in advance by the mechanics of the world, to the capricious, unrestrained free will that puts intention and responsibility at each individuals’ complete disposition and initiative. The truth lies probably somewhere in between. There are likely neuronal mechanisms involving stochasticity that are constrained by external influences and generate choice options within a limited range. Deliberate actions might originate from these neuronal mechanisms and operate within these constraints. Constraints are provided by reward value and all factors influencing it, including effort cost, delay, risk, and references, that affect brain processes underlying decision-making. Without trying to provide a coherent view, I will sketch a few thoughts along these lines.

1. Origin of free will (the impossible question)

Assuming that free will exists at all, we simply do not know how it may originate, and we have only faint ideas how brain processes might generate it. We argue even whether free will originates in the brain. For Kant, free will lies above all activity, irrespective of physical, biological, psychological, and social constraints. For him, only God has free will, as s/he has created everything. But admitting that we do not know anything is the easy answer. We can try to reduce this impossible question to tractable issues by trying to understand the mechanisms determining free will.

Free will has several necessary components. First, free will depends on the ability to make choices, to do what we want to do (219). Second, free will implies a voluntary action that is self-directed and arises on self-command. Free will does not apply to vegetative reactions that I cannot control, like heartbeat, intestinal contractions, or hormone release, nor to probably most Pavlovian reactions and many habits. Third, the word will indicates a goal that we want to achieve. Thus free will would be instrumental in generating voluntary, goal-directed behavior. Fourth, free will involves conscious awareness about the goal directedness. We think we have free will because we know explicitly that we are working towards something we want. And we know that we want this something, however diffuse this may be in exploratory behavior. We feel the freedom to choose when to act and which of several options to select. This feeling is subjective, private, and unobservable by others. It is difficult to measure objectively, and it may not even be real but an illusion (634), nobody knows for sure.

The fact that free will has emerged in evolution suggests a beneficial function for the competitive survival and reproduction of individuals in their function as gene carriers. As much as the idea of a conscious illusion of free will may be disappointing, evolutionary biology teaches us that a structure or function may enhance the chance for survival without necessarily evolving to perfection. Thus it may not matter whether conscious free will is an illusion or real, as long as it is helpful for keeping the genes in the pool. Thus free will is a psychological tool for gene propagation.

What we do know is that we need rewards for survival (essential substances contained in food and drinks) and reproduction (sexual rewards). It would be advantageous to seek these hidden rewards actively rather than waiting for them to appear out of the blue. That is why we can move. Intentions and free will would be a great help to seek and explore rewards actively and deliberately, rather than running after them using deterministic programs. Without the true or false belief of free will, we may have only limited initiative to find hidden rewards. The main advantage of free will is the subjective experience of free will, which makes the search for hidden rewards more deliberate and focused, and thus saves time and is more efficient in many situations. Dennett (128) argues that consciously sensing free will is a crucial characteristic of free will. The evolutionary benefits driving the appearance of free will would be the boost in initiative, resulting from the conscious belief that we can deliberately decide our own behavior. This belief is also the basis for moral responsibility and thus helps social coherence. And adding theory of mind would make me understand others’ free will and voluntary behavior and further facilitate social processes.

One view on conscious free will assumes that it is generated from scratch and makes us completely free to do what we want. There are only two possibilities to generate this sort of free will. First, brain activity generates the conscious awareness as necessary hallmark of free will. This brain activity necessarily precedes the conscious awareness and thus cannot be directed by free will. In this case, free will would not be free but an illusion of freedom that follows the uncontrolled generation of free will.
by the brain. Alternatively, it is not brain activity but something nonmaterial that leads to conscious awareness of free will, and then brain activity carries out what the nonmaterial event has directed it to do and signals free will. Both scenarios seem unresolvable and therefore depressing and ultimately unconstructive. However, the problem may be overcome by taking a step back to consider the neuronal mechanisms by which free will might arise in more realistic situations in which it is seriously constrained by a number of factors. Even in this restricted form, free will would still provide an evolutionary beneficial, conscious control over our actions.

Everyday neurophysiological experiments tell us that many, maybe most, neurons in the brain are spontaneously active without any observable stimuli or actions. Thus brains generate their own activity. We also know that we have many mental experiences that seem to arise spontaneously out of nowhere, even in fully awake and nonpathological states, like thoughts, imagination, and old memories. How could I write this text if I could not sit back (with the proverbial cup of tea) and listen to the ideas that are popping up in my head (most of them rubbish, but occasionally a useful one), and then find the words and put them into grammatically correct order? Neuronal activity in cortex and striatum ramps up over several seconds before self-initiated movements but not before externally instructed actions (299, 311, 337, 380, 397, 491, 493, 526, 528). The key to explaining such brain activity may be the tendency to stochastic activity in many neurons and networks (129, 167, 180, 195). Once initiated by random molecular fluctuations (94, 287, 561, 567), the noise may result at some point in organized activity (168, 200, 297, 330, 465, 564, 611) and influence or settle in attractor states (63, 125, 510) that result in mental events. The hardware of the individual’s brain and the individual’s experience would constrain the range of attractor states producing mental events. A person who has never done or watched or heard about underwater diving may never experience a spontaneous imagination of the color of corals, nor the free will to go diving to see this. Thus the stochastic brain activity would arise spontaneously, relate to past experiences, and not be initiated from scratch by free will (nor do we need to assume nonmaterial origins).

The spontaneously occurring attractor states arising from initially stochastic neuronal activity would not need to be conscious. Several such elementary unconscious attractor states may develop at the same time, each of them representing a different future option for intention. The possibility of developing attractor states from stochastic activity would not be infinite but depend on brain activity restricted by hardware, current environment, and memory from earlier experience. Thus the attractor states would be restricted to viable options within boundaries. Impulses that never come up cannot be selected. Attractor states related to the intention of exploring the color of corals may only occur if my brain can process color and if I have ever heard about colorful corals. Then a selection mechanism would be set in motion in which viable options emerge with a certain degree of variation within bounds defined by hardware and personal experience. The process may resemble the selection of a reasonable option from a rotating initiator of options. There would be no deliberate, free initiative involved in this selection, as not only the attractor states but also the selection process and its boundaries would be unconsciously mechanical. Only the result of the selection would pop up into consciousness and be interpreted as free will. This conscious free will is not an illusion but the result of unconscious processing. From this point on, it would be possible to decide consciously whether to follow the selected impulse for action or to resist, to select consciously between the emerging mental images and intentions (e.g., separating rubbish from non-rubbish in writing), or to listen further to solutions from those popping up. Thus the unconscious processes produce a scratchpad that pops into consciousness and from which conscious free will can choose freely or veto an intention. This might be what is usually associated with the term of free will and the individual’s responsibility for their own intentional actions.

Of course, there may be completely different, currently unknown processes besides stochasticity that may generate brain states underlying free will. The idea would require the discovery of new phenomena in physics (429). However, some of the described steps in the emergence of free will may still hold, in particular the dilemma in the initiation of free will between the brain’s hardware and the mind.

2. Constraints on free will

Although free will is associated with being able to make choices (219), I can only choose between the options that are available. We are just not free to choose between anything imaginable, but our options are restricted by the physical state of the world (we cannot fly to Mars), our body’s hardware (we cannot fly like birds), and our personal experiences and education (most language scholars do not do partial differential equations). Even among the available options, our choices are influenced by factors not considered as economic value such as emotions driven by external events, social pressure and conventions, altruism, moral rules, and irrational beliefs such as nationalism, chauvinism, prejudices, and superstitions. These external and internal constraints reduce free will in a graded manner, and up to its complete loss when constraints become too serious. Of course, survival is the strongest and most natural constraint on free will. Only within these bounds is free will free.
Free will can be constrained and focused. A good example is intention in action (529). I choose deliberately a course of action that allows me to form short-lived intentions on the fly while I am moving. I choose to do a walk through town and find a shop window that attracts my intention and makes me buy a pair of boots for the winter. My free will is partly restricted by the decision to take the walk, and the shoe purchase reflects that constraint. My purchase may simply be triggered by the view of the shoes, their price, and their relationship to my purchasing habits and personal preferences. The intention to enter the shop comes instantaneously, almost as a reaction to the sensory input. Very little free will is involved in the final purchasing action after I have initiated the walk.

In a similar situation of very restricted free will, humans indicate the moment at which they consciously perceive the awareness to initiate a simple finger movement (FIGURE 49A) (316). Indeed, the participants become aware of the urge to do a self-initiated finger or wrist movement and report the dot position after being prompted. B: electroencephalographic readiness potential recorded at the vertex (top of skull) of one human participant [average of 40 trials]. Onset of aware urge to move is indicated in red (bar shows range from eight averages of 40 trials each). Muscle onset denotes onset of electromyographic activity in involved forearm muscle. [From Libet et al. (316), by permission of Oxford University Press.] C: detection of time and occurrence of deviation from baseline activity in 37 single and multiple neurons before awareness to act, as determined by a support vector machine classifier. The neurons or multineuron clusters are located in human supplementary motor area (n = 22), anterior cingulate cortex (n = 8), and medial temporal lobe (n = 7). D: ramping activity preceding the conscious awareness to act in activity averaged from 59 neurons in human supplementary motor area. [C and D from Fried et al. (172), with permission from Elsevier.]
as if the brain’s hardware initiated the movement without the participant’s own initiative, suggesting that conscious free will does not exist in these experiments. It is tempting to deny the general existence of free will on the basis of these results. However, the voluntary submission to the experiment may put so many constraints on the liberty of the participant that indeed there is no free will left and the brain is able to initiate the movement without conscious free will, just like the shoe buying example. Thus the experiments employ the role of constraints on free will in an extreme manner and test intention in action rather than true free will. In doing so they may reveal the existence of the elementary process outlined above by which stochastic brain activity settles spontaneously into an attractor state without conscious awareness and only afterwards leads to a conscious mental state. The consciously aware urge to move is then simply a retrograde interpretation of an unconsciously initiated action following the conscious selection to undergo the experiment.

3. Rewards exert constraints on free will

Rewards are necessary for survival. Without rewards we would die of thirst within days and of hunger within weeks, depending on our adipose reserves. With that function, rewards influence and thus constrain behavior and restrict free will in several ways. The passive experience of rewards attracts behavioral reactions away from other objects and focuses behavior onto the available rewards. Entering active decisions between rewards is not very free either because we need rewards at some point for survival. Once we are able to decide, we choose the best option, called utility maximization in economics, to ensure the highest chance for survival and gene propagation. Choices in such restricted situations may often reflect intentions in action, similar to the Libet experiments on movements, and involve serious constraints on free will. The choices are easy to do when the options are limited and have clearly different values, in which case there is little scope for free will, but the choices are more difficult when more, and more similar, options remove some constraints on free will. Habits for acquiring rewards impose particularly strong constraints on free will, and addictions to natural and artificial rewards eliminate free will even more. The initial administration of an addictive substance, like nicotine, is done with free will, often constrained by peer pressure, but following the habit once serious addiction has set in is almost deterministic.

All additional factors affecting reward value add to the constraints on free will. Motivational states like hunger and thirst make rewards more valuable and prioritize our behavior, as easily seen before lunch and dinner. Rewarding contexts make us long for rewards. Who would work during an office party when colleagues drink wine and tell interesting stories? Work required to obtain rewards seriously focuses our actions. Many animals spend most of their day searching for food. Temporal discounting of future rewards hinders our planning and favors myopic actions. Reference-dependent valuation makes us slaves of our environment. Nobody wants to lose what they have achieved, and keeping up with the Joneses is hard work, both of which require good focus and constitute constraints.

Risk exerts huge constrains on our behavior. Risk avoidance makes us shy away from risky situations even if they might be beneficial in the long run. The inability to completely eliminate risk induces anxiety, prejudice, and superstition which further direct behavior in serious ways, up to generating wars without reason. Risk management and the desire for risk reduction are prevalent in all cultures. Many activities of daily life are aimed at reducing risk, from paying car and health insurance to ritual sacrifices. The first reaction of the Minoan people to the devastating Santorini volcano eruption was human sacrifices to appease the angry gods. Risk reduction becomes a part of the culture, influences the whole life history of individuals, and thus takes some free will away.

The constraints exerted by rewards on free will extend also to the learning about rewards. Pavlovian reward predictions arise automatically following repeated exposure without the subject’s own doing, which constrains free will substantially. Operant conditioning involves own action and allows goal-directed behavior, which constrains free will a bit less than Pavlovian predictions. Conditioning and higher forms of learning are evolutionary beneficial, as they increase reward acquisition for survival and gene propagation, but in doing so they constrain free will considerably.

Genetic dispositions, education, and long-term reward experience are instrumental for establishing personal reward preferences that influence the whole life of individuals. Different individuals prefer different rewards and choose their profession accordingly (among many other factors). They become mechanics because they appreciate precision, bankers for the money, musicians for the beautiful sounds, doctors because of altruism, and scientists because they value curiosity and knowledge highly. The impact of these preferences accumulates over the whole life span and shapes the personality. All of these influences constrain free will. Even seemingly free changes in personal reward preferences due to physiological factors and personal experience do not much enhance free will.

Punishers constitute the second major class of motivators of behavior besides rewards. Punishers derive from external sources over which we have little control, and they need to be avoided to escape their damaging impact. Thus avoidance is a reaction to events that occurs without one’s own doing. There may be different options for escape, and thus limited scope for free will, but overall avoidance involves little free will, allows little deliberation, and produces primarily reactive rather than proactive behavior. The choices
between punishers seem to allow some degree of liberty that is intrinsically low when being forced into the choice. According to the New York Times of April 23, 2010, prisoners sentenced to capital punishment in Utah may choose between hanging and a firing squad. “I would like the firing squad, please” reportedly said Ronnie Lee Gardner. Overall, some liberty exists when being able to proactively search for best solutions for efficient avoidance. We can think of known dangers and generate preventive measures, but it is hard to imagine totally unknown dangers and be creative in planning their avoidance. Thus punishers may exert even more constraints on free will than rewards.

4. Neuronal reward processing constrains free will

Electrical and optogenetic stimulation of dopamine neurons are very efficient for eliciting immediate behavioral reactions and behavioral learning (FIGURE 19, A–D) (2, 103, 155, 562, 605, 641). The strength of dopamine stimulation affects choices, reaction times, and reward acquisition in a graded manner. The animals are simply driven towards goals associated with the dopamine activation, both during the stimulation (immediate effect) and in the next test sessions (reinforcement learning). Thus, in focusing individuals on rewards at the expense of other pursuits, dopamine stimulation takes parts of free will away from the individual’s behavior.

Dopamine stimulations are not experimental artifacts but mimic the natural phasic dopamine responses to reward prediction errors. The dopamine responses are likely to have similar behavioral effects as the stimulation. Even though stimulation parameters differ somewhat from the natural responses, the important feature is the brief, phasic dopamine activation, whereas slow, tonic dopamine stimulation is ineffective (e.g., for place preference learning; Ref. 605). The dopamine effects likely involve synaptic plasticity and focusing in striatum and frontal cortex (70, 368, 424, 605). Indeed, optogenetic stimulation of striatal neurons postsynaptic to dopamine neurons induces differential approach and avoidance learning and behavioral bias depending on the dopamine receptor subtype the stimulated neurons carry (FIGURES 19F AND 46C) (293, 576). Thus the combined evidence from dopamine stimulation and dopamine reward responses indicates that eliciting a strong reward signal in dopamine neurons directs the organism towards reward and thus constrains its free will. This is the mechanism through which primary (unpredicted) rewards and (Pavlovian) reward predictors influence the behavior and constrain free will.

A similar argument may hold for neuronal risk processing. Dopamine and orbitofrontal reward neurons code risk separately from value (FIGURE 31C) (161, 162, 391, 392). Furthermore, neuronal risk signals, and the influence of risk on neuronal value signals, vary with individual risk attitudes (301). If these risk signals could affect risk components in behavior analogous to value signals, they would provide mechanistic explanations for the constraining influence of risk on free will. To demonstrate such an effect, one would need to demonstrate that selective stimulation of risk neurons, such as those found in orbitofrontal cortex, affects risk attitudes during choices.

The strength of the dopamine response seems to convey the influence of rewards on behavior. Rewards with stronger effects on dopamine neurons are likely to have more impact on learning and choices. Thus the propensity of individual rewards to activate dopamine neurons would determine the influence of reward on behavioral choices. This function may hold also for other reward centers in the brain, although their less compact organization may not be conducive to such dramatic stimulation effects. However, it is unlikely that a response of dopamine neurons or other reward neurons to an automatically conditioned, reward-predicting stimulus induces approach behavior entirely without the subject’s own doing. There are other brain mechanisms that would limit the automaticity of such effects, but any neuronal reward signal is nevertheless likely to have a biasing effect towards rewarded stimuli and actions. The neuronal responsiveness to different rewards is likely to vary between individuals. Interindividual differences in reward processing may affect daily preferences but also determine long-term behavior, including professional choices and other important decisions in life. Thus the activity of reward neurons shapes behavior, constrains voluntary decisions, and thus restricts free will.

ACKNOWLEDGMENTS

I thank Anthony Dickinson for 20 years of discussion and collaboration on animal learning theory, Peter Bossaerts and Christopher Harris for 10 years of discussion and collaboration on experimental economics, and Peter Bossaerts, Anthony Dickinson, Fabian Grabenhorst, Armin Lak, Johannes Schultz, William R. Stauder and two anonymous referees for comments. I am indebted to Colin F. Camerer, David M. Grether, John O. Ledyard, Charles R. Plott, and Antonio Rangel for discussing experimental economics at the Division of Humanities and Social Sciences, California Institute of Technology, Pasadena.

Address for reprint requests and other correspondence: W. Schultz, Dept. of Physiology, Development and Neuroscience, Univ. of Cambridge, Cambridge CB2 3DY, UK (e-mail: ws234@cam.ac.uk).

GRANTS

This work is supported by Wellcome Trust Principal Research Fellowship, Programme and Project Grants 058365, 093270, and 095495; European Research Council Advanced Grant 293549; and National Institutes of Health Caltech Conte Center Grant P50MH094258. Previous sup-
port came from the Human Frontiers Science Program, the Wellcome-MRC-funded Behavioural and Clinical Neuroscience Institute Cambridge, and the Swiss National Science Foundation.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author.

REFERENCES


Physiol Rev © VOL 95 • JULY 2015 • www.prv.org

NEURONAL REWARD AND DECISION SIGNALS


433. Platt ML, Glimcher PW. Neural correlates of decision variables in parietal cortex. 


484. Roitman MF, Wheeler RA, Carelli RM. Nucleus accumbens neurons are innately tuned for rewarding and aversive taste stimuli, encode their predictors, and are linked to motor output. *Neuron* 45: 587–597, 2005.


NEURONAL REWARD AND DECISION SIGNALS


