Many decisions are based on past experience because we expect the values of choices to be stable or to change only gradually. To choose effectively, we must either track these values or, ideally, anticipate the future value of a choice. While much is known about how choice values are learned from past experience, it is relatively unknown how people predict in advance not only that a value will change but also how it will change. This problem lies at the heart of many judgment tasks, such as stock market trading or behavioural adaptation to the continuously changing reward rates animals and humans experience in foraging patches[1,2]. A simple way, however, to predict future choice values is to estimate the change in rewards. This can be performed by comparing recent and past reward rates.

Evidence comparison is central to decision making[3] and value comparison is central to reward-guided decision making. Ventromedial prefrontal cortex (vmPFC) signals reflect the comparison of choice values when making decisions[4]. VmPFC blood oxygen-level-dependent (BOLD) activity scales with the value difference between available options, and the chosen and unchosen options have dissociable positive and negative effects[4–7], which can be understood as reflecting a value competition process[8–11]. Previous studies focused on decisions between two clearly defined and concrete options usually associated with specific stimuli. While this is clearly an important decision-making mode, behaviour often also reflects alternation between continuous engagement in the same behaviour (for example, when an animal repeatedly forages in one patch or a person maintains engagement with the same task) with behavioural change (when the animal moves to an alternative patch or the person switches task)[1,2,12,13]. In such situations it is critical to know prospectively how profitable the next repetition of an action will be. For some time there has been evidence that evaluation of a choice reflects comparison against previous choice values. For instance, rats' approach speed to a given reference amount of food pellets depends on their previous reward experience. Having experienced more rewards previously, rats are slower to approach the reference amount of food pellets; contrarily, having experienced fewer rewards beforehand, they are quicker to approach the same amount of food. These phenomena are known as positive and negative successive contrast effects, respectively[14,15]. Similarly, deciding whether to further commit to, or to quit, engagement in an action, the critical comparison is not between two stimulus values but between elements of the past reward history.

Brain mechanisms comparing recent and past reward rates require fine-grained information about the past history of rewards. Dorsal anterior cingulate cortex (dACC) is necessary for retaining the history of encountered rewards[16,17], and neurophysiological studies have demonstrated dissociable influences of past and recent rewards on the activity of dACC neurons[18,19]. In addition, activity in the dopaminergic system and striatum also reflects reward history[20–22]. Reward history signals here have been understood in the context of reinforcement-learning models[23] that involve computation of a reward prediction error (PE) relative to previous outcomes[24]. However, typically, simple reinforcement learners cannot weight recent and past rewards in an opposing manner. To do this, a learning mechanism requires a contrast mechanism comparing recent and past reward rates[15,25,26]. We used functional magnetic resonance imaging (fMRI) in humans to investigate value comparisons of recent and past reward rates. We find that dACC holds multiple time-linked reward representations simultaneously, predictive of the way past and recent rewards guide decisions to stay or to leave an environment. Our results suggest a key role for dACC in computation of reward trajectories and the transformation of decision variables to choice.

Results

Experimental design. We designed a reward-learning task in which subjects chose to further commit to, or to leave, a foraging-like patch based on its estimated future value. The patches were characterized by reward rate trends that could be discerned by comparing past and recent reward rates. If the decision maker knows the reward rates at two different time points in the past, they have sufficient information to judge whether the reward rate has increased or decreased between these time points. In environments with monotonic reward rate changes such knowledge can be exploited to extrapolate the reward trend and predict the future value of the patch. Patches were derived from reward rate curves similar to those in optimal foraging theory[2] (Fig. 1a,b and Supplementary Fig. 1). The patches consisted of sequences of time steps on which either reward or non-reward events occurred (Fig. 1c,d). The reward events were spread out such that their reward rates conformed to the underlying reward rate curves. Subjects proceeded from time step to time step by pressing a button. At a predetermined time step, subjects were offered a leave–stay decision (LSD; Fig. 1e). For LSDs, subjects had to consider the 15 further time steps they would encounter after LSD, and decide, for this time period, whether to stay and further explore the environment they were in or to leave and re-engage with a pre-learned default environment with a stable reward rate. LSDs should be based on a comparison of the anticipated value of the current environment (the sum of rewards that would be delivered after LSD) and the pre-learned value of the default environment.

Opposing effects of recent and past rewards on choice. We measured the influence of rewards occurring at different times during a trial on the LSD using a general linear model (GLM). Therefore, we divided the reward history into reward rate bins reflecting reward received over five sets of three time steps each moving backwards in time from the LSD (LSD-1-3, LSD-4-6, and so on; Fig. 2a). In our experiment, reward rates in these different time bins share less than 25% of their variance; therefore, we can estimate their influences on choice behaviour. Note also that this means that we could also test whether choices were solely based on a patch's initial reward rates or whether later reward rates had an additional influence on choice (correlation between initial and last time bin: $r=0.02$), which turned out to be the case. While subjects tended to stay in a patch when reward rates in recent bins were high (LSD-1-3: $t_{19}=10.12$; $P=4\times10^{-9}$; LSD-4-6: $t_{19}=4.98$; $P=8\times10^{-5}$), the more distant the reward rate bins were the more negative the effect of high reward rates on the decision to stay (LSD-10-12: $t_{19}=-6.85$; $P=2\times10^{-6}$; LSD-13-15: $t_{19}=-6.99$; $P=10^{-6}$).

We compared subjects' choices with choices made by an individually fitted simple Rescorla–Wagner reinforcement-learning (RL-simple) model[23]. We used the value estimate of the model at the time of the LSD (that is, after observation of the last reward outcome) as decision variable. We applied the former GLM to RL-simple-simulated choices (Fig. 2a, green bars). RL-simple captures recent positive effects of rewards (LSD-1-3: $t_{19}=6.44$; $P=4\times10^{-6}$; LSD-4-6: $t_{19}=7.58$; $P=4\times10^{-7}$), but is unable to simulate a negative influence of past reward rate bins on choices to the same degree as seen in human subjects (paired t-tests on last two bins: LSD-10-12: $t_{19}=8.11$ $P=1\times10^{-7}$; LSD-13-15: $t_{19}=6$; $P=9\times10^{-4}$). The strongly negative effects of rewards in past time bins in our human subjects (LSD-10-12, LSD-13-15) resemble successive contrast effects in animals[14] (inset Fig. 2a).

The temporal gradient in Fig. 2a can be summarized using two parameters: the reward rate of the last reward event (lastRR)

before the LSD and the average reward rate (avgRR) throughout the whole period before the LSD (Fig. 2b). These two regressors capture the negative and positive effects of past and recent rewards on human behaviour (lastRR: $t_{19}=9.56$; $P=10^{-7}$; avgRR: $t_{19}=-6.23$; $P=5\times10^{-6}$). While the RL-simple model explained the lastRR effect ($t_{19}=7.78$; $P=3\times10^{-7}$), it falsely predicted a positive instead of a negative effect for avgRR ($t_{19}=3.53$; $P=0.0023$; see Supplementary Fig. 2a,b for a description of the complementarity of Fig. 2a,b). The difference between lastRR and avgRR (lastRR − avgRR) gives a measure of reward rate trend, accounting for both increasing (positive) and decreasing (negative) reward rate trends (Fig. 2c and Supplementary Fig. 2c,d). As our investigations focused on the use of any type of monotonic reward trend in choice, lastRR − avgRR assumes the simplest type of trend, a linear one, rather than, for example, an exponential trend. Note, however, that subjects based their choices not only on lastRR − avgRR, but also on the absolute size of lastRR (Supplementary Fig. 2c).

The shortcomings of RL-simple become particularly clear when binning the trials by their categorical reward rate trend (increasing or decreasing) and sorting them by optimal choices (defined by maximal payoff; Fig. 2d). When choices had to be made against the reward rate trend (decreasing/stay-optimal and increasing/leave-optimal), RL-simple's choice predictions were close to the choices observed in subjects (although for decreasing/stay-optimal trials, we found a small but significant difference $t_{19}=2.44$; $P=0.025$). In these cases, correct choice could be based on lastRR alone. However, when optimal decisions depended on the reward rate trend (increasing/stay-optimal and decreasing/leave-optimal), human behaviour was strikingly more optimal than that predicted by RL-simple ($t_{19}=9.72$; $P=8\times10^{-9}$ and $t_{19}=-10.31$; $P=3\times10^{-9}$, respectively). These results corroborate earlier results (Supplementary Fig. 2c) that subjects base their choices on both the instantaneous reward rate at the time of choice and the reward rate trend.

In sum, we show that subjects' choices were influenced by recent and past reward rates in an opposing manner. Because a simple-RL model integrates historical and recent rewards into a single estimate of an option's value, it does not represent recent and historical rewards with opposing weights and, therefore, does not make choices like the human subjects.

Estimating PEs enables trend-guided choice. Prediction errors (PEs; differences between the reward outcomes that choices actually led to and prior reward expectations) are a simple measure of the option's reward rate change. Because PEs have a role in value-updating they have been linked to learning[23]; however, it is also possible that PEs may be used as a decision variable to guide decisions[25]. We tested whether an RL model

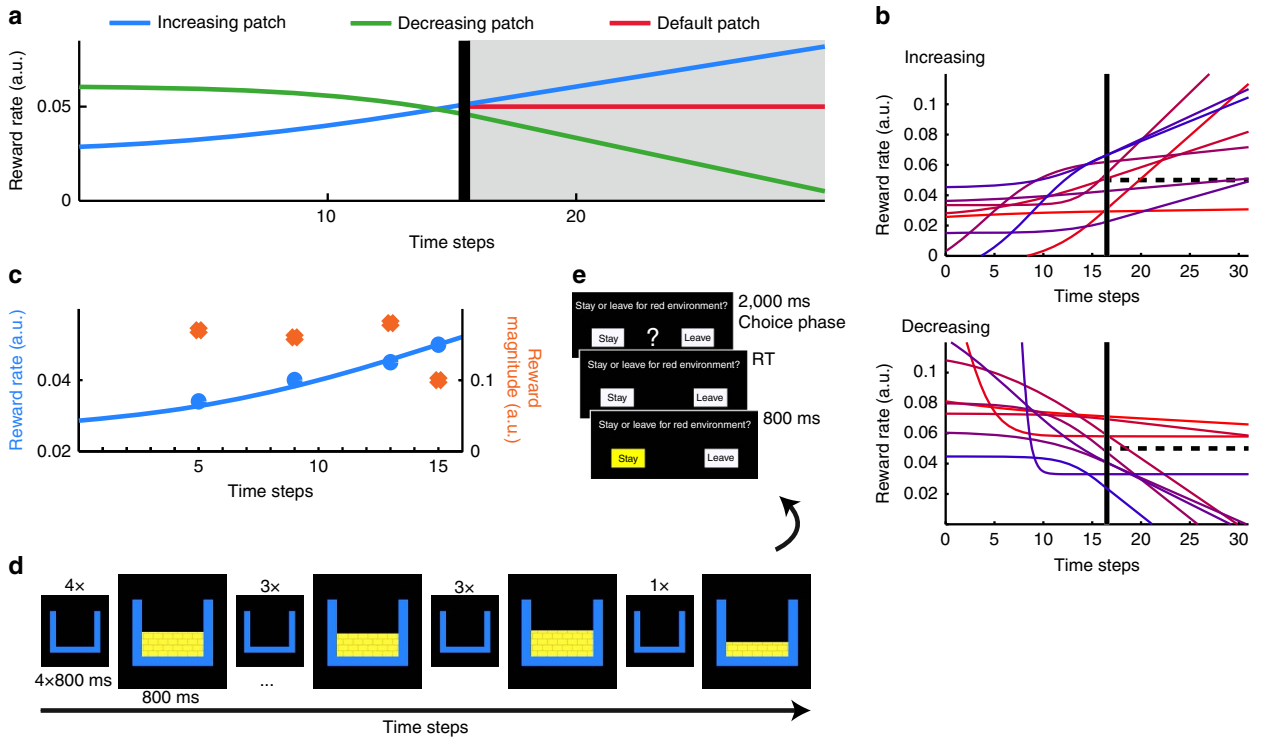

Figure 1 | Experimental design and implementation. (**a**) Two example patches with increasing (blue) and decreasing (green) reward rates. At LSDs (black) subjects chose between staying in the patch and switching to a default patch with a stable reward rate (red). In these examples, correct decisions (stay on blue, leave green) can be predicted from the reward rate curves. (**b**) All 18 reward rate curves from which trials were derived. Solid black line indicates LSD and vertical dashed line indicates reward rate of reference patch. For visualization purposes only, different colouring for reward rate curves was used, and the curves were aligned so that the LSD is on the same time step. (**c**) Sequence of events corresponding to the blue reward rate curve before LSD in **a**. Four reward events were presented at time steps 5, 9, 13 and 15. Their reward rates (blue dots), which conform to the reward rate curve (blue line), are calculated by dividing their reward magnitudes (orange dots) by the time delay from the previous reward event or the start of the patch. (**d**) Screen during events in **c**. Empty boxes represent non-reward events; the height of 'gold bars' in reward events represent their reward magnitudes. Each event was displayed for 800 ms. Between events, a fixation cross was shown and subjects proceeded to the next event by pressing a button. Note that lastRR for this patch would be equivalent to the height of the gold bars in the final box divided by two time steps. (**e**) LSD followed, without time jitter, after the last reward event (boxes in the default environment were red; therefore, it was labelled 'red environment' for subjects).

prefrontal cortex[50,51]. The finding that dACC activity was related to both reinforcement history-derived variables and ensuing choices further supports the contention this area represents and translates reward history representation into a choice-related representation, as had also been suggested by the correlation between dACC neural and behavioural beta weights (Fig. 4c,d,f). In both analyses as well as in previous reports[28–30], choice-related representations were present in dACC beyond difficulty signals. Moreover, choice-related signals in vmPFC and dACC were aligned in this experiment; both areas are more active for stay compared with leave choices and dACC activity increases with evidence for staying. In previous experiments, vmPFC activity was positively related to the value of staying versus discarding a current choice, while dACC was positively related to exploring versus exploiting a choice[28,34,52–54]. However, while the value of the stay-option in the previous experiments was relatively well known and stable, it had to be inferred in the current experiment because the future value of a patch is, by design, different from its past value. The leave-option was, on the other hand, pre-learned and stable. In other words, the outcome uncertainty for the stay-option is, unusually, higher than that for the leave-option. One hypothesis that integrates the present results with previous findings is that dACC codes value in a framework tied to pursuing the more uncertain option; by contrast, vmPFC may code the value of staying with the current or default choice[28,34,55,56]. The two reference frames are aligned in the current experiment but have been opposed in previous experiments.

Methods

Subjects. Twenty-two subjects participated. One left the experiment as a result of claustrophobia. Another was excluded from data analysis due to excessive motion (final sample: 20 subjects; eight female; aged 21–32). All provided informed consent. The study was approved by the Ethics Committee of Oxford University (MSD-IDREC-C1-2013-095). Subjects received £20 as a show-up fee and a fraction of £15 depending on task performance.

Experimental design. In each of 90 trials, subjects proceeded through a patch consisting of reward and non-reward events (800 ms each). Between events, a fixation cross was presented and button presses led to the next event. After time step 15, 16 or 17, the subjects were offered the choice to stay for longer in the patch or to leave to a default environment with a known, stable reward rate (leave-stay decision; LSD). This meant that participants encountered LSDs at approximately similar positions in time on each occasion that they explored a new environment; however, the presence of some variability in LSD timing precluded precise anticipation of the LSD time by participants. Moreover, the analyses do not focus on activity that is simply linked to the main effect of LSD occurrence but to activity that is parametrically related to the reward experience before the LSD and that allowed the subjects to make inferences about the reward likely to be received after the LSD. The design of the trials was based on 18 monotonic reward rate curves, nine increasing and nine decreasing, from each of which five unique sequences of reward/non-reward events were derived (Fig. 1). The key manipulation was to assemble a set of patches such that the behavioural and neural effects of recent and past reward rates could be dissociated. The reward rate of a reward event was its reward magnitude divided by the number of time steps from the previous reward event or from the start of the sequence (time delay; Fig. 1c,d). The reward magnitude was indicated by the height of a golden texture ('gold bars') within a box presented on the screen. The time delay between reward events ranged between two to six time steps. On every trial, the LSD was, without time jitter, preceded by a reward event at the last time step of the sequence to keep the recency of the last reward event with respect to the LSD constant. Note that this meant that the last time step was always rewarded and the second last time step was never rewarded. Each LSD began with a 2,000 ms 'choice phase' indicated by a question mark on screen (referred to as 'LSD' in time-course plots). Subjects then responded and their choices were highlighted in yellow for 800 ms (Fig. 1e). The left-right locations of 'stay' and 'leave' buttons were randomized. After LSDs, the subjects continued through a sequence of 15 additional time steps. For stay choices, the underlying previous reward curve continued linearly based on its slope at the time of the LSD. If subjects decided to leave to the default environment, rewards were delivered at a fixed reference reward rate (that is, the ratio of reward magnitude and time delay was identical for all reward events of the sequence) that subjects knew well from prior practice sessions. Although the specific event sequence differed between runs of the default environment, the sum of the rewards encountered there (leave value) was constant. For LSDs, subjects had to decide whether the future value of the patch, the sum of reward encountered when further committing to it (stay value), was higher or lower than the stable leave value. On 66% of increasing patches, the stay value was higher than the leave value, while on 66% of decreasing patches, the leave value was higher than the stay value. Subjects received 'bonus points' for making the better choice in proportion to the absolute difference between the stay and leave values and this determined the performance-dependent monetary payoff subjects received at the end of the fMRI session. Optimal choices were defined as the ones with the higher payoff. Subjects received feedback about their accumulated bonus points after 1/4, 2/4 and 3/4 of the experiment. Finally, 60% of the trials were truncated after the LSD to shorten the experiment and to minimize choice feedback.

Reward sequences. The generation of actual reward sequences from the theoretical reward rate curves was an iterative process. Beginning at the last time point of the curve (because the last event before a LSD or the end of a patch was always a reward event), a reward magnitude and a reward delay were randomly chosen under the condition that they conformed to the reward rate indicated by the reward rate curve at that time point. The chosen reward delay determined the time point of the reward event preceding this event. For this, again, reward magnitude and reward delay was randomly chosen under the condition that it they conformed to the reward rate indicated by the reward rate curve at that time point. The threshold for 'conforming to the reward rate curve' was initially set to 5% of the reward rate given by the curve; however, it was expanded in 0.00001% steps in case no solution could be found. Note that for our analyses, it did not matter how closely the reward rates of the reward events in the sequence satisfied the reward rates originally intended for the curves because all parameters used in our analysis were derived from the actual reward sequences and not from the reward rate curves. In other words, the reward rate curves that we show and use to guide all analyses of behaviour and neural activity are the ones that were established empirically at the end of this iterative process.

Training session. Subjects experienced a 60–70 min training session on the day before scanning and a 15–20 min training period directly before the scan. Both training sessions comprised familiarization with the default environment and a version of the experimental task with choice feedback (number of bonus points earned or missed) after each trial (Supplementary Fig. 1).

Reinforcement-learning models. To compare subjects' choices with a learning algorithm that integrates reward rates with a single time constant we devised a standard Rescorla-Wagner reinforcement learning model (RL-simple)[23]. More complex models are based on RL-simple (see below).

At the beginning of a patch, a value estimate was set to the average reward rate experienced in the task up to that point and was zero for the first trial. The value estimate was then updated by the outcome of each time step of the patch sequence. Therefore, the number of value updates in a patch before LSD was equivalent to the number of time steps in the sequence. The value estimate was updated using a learning rate α fitted for every subject:

$$\mathrm{Value}(t+1) = \mathrm{value}(t) + \alpha \times (\mathrm{outcome}(t) - \mathrm{value}(t)) \quad (1)$$

The size of the outcome was zero or positive, and reflected the reward magnitude encountered at a time step. The value estimate at the time of the LSD ($t = \mathrm{LSD}$), that is, after the last event of the sequence, was used as the decision variable (DV):

$$\mathrm{DV} = \mathrm{value}(\mathrm{LSD}) \quad (2)$$

The DV, representing the value of staying in a patch, was compared with the value of the default patch to determine the model's choice. The probability of staying in a patch was calculated with a softmax equation:

$$P(\mathrm{stay}) = \frac{\exp(\beta \times \mathrm{DV})}{\exp(\beta \times \mathrm{DV}) + \exp(\beta \times \mathrm{value_{DEF}})} \quad (3)$$

β is the inverse temperature of the softmax function and the constant $\mathrm{value_{DEF}}$ represents the value of the default environment. Note that modelling $\mathrm{value_{DEF}}$ as a free parameter means we do not use a RL mechanism to learn the value of the reference patch. Given that the reference patch is pre-learned and only very rarely encountered in the actual experiment (60% of trials end directly after the LSD), we determine the value each subject assigns to the reference patch empirically by treating $\mathrm{value_{DEF}}$ as a free parameter that is stable over the course of the experiment. $\mathrm{Value_{DEF}}$ was used in all RL models. We derived the choice probability from the stay probability on each trial:

$$P(\mathrm{choice}) = \begin{cases} P(\mathrm{stay}) \text{ if stay} \\ 1 - P(\mathrm{stay}) \text{ if leave} \end{cases} \quad (4)$$

Overall, the free parameter set θ comprised α, β and $\mathrm{value_{DEF}}$. We fitted these parameters for every subject by minimizing the negative log likelihood (nLL) over