TIME ELAPSED BETWEEN CHOICES IN A PROBABILISTIC TASK CORRELATES WITH REPEATING THE SAME DECISION

Judyta Jabłońska[1], Łukasz Szumiec[1], Piotr Zieliński[2], Jan Rodriguez Parkitna[1]

[1]Department of Molecular Neuropharmacology, Institute of Pharmacology, Polish Academy of Sciences, Krakow, Poland

[2]Department of Structure Research of Condensed Matter, The Henryk Niewodniczański Institute of Nuclear Physics, Polish Academy of Sciences, Krakow, Poland

Corresponding address:

Jan Rodriguez Parkitna,

Jan.Rodriguez@if-pan.krakow.pl,

Department of Molecular Neuropharmacology,

Institute of Pharmacology of the Polish Academy of Sciences,

Smętna 12, 31-343 Krakow, Poland

Running title: Effect of time on probabilistic choice

27 pages, 8 figures, 1 table, and 12 equations

The total number of words in the whole manuscript including this page: 7826

The total number of words in the abstract: 237

Key words:

Reinforcement, mice, alcohol, reward, learning

ABSTRACT

Reinforcement learning makes an action that yielded a positive outcome more likely to be taken in the future. Here, we investigate how the time elapsed from an action affects subsequent decisions. Groups of C57BL6/J mice were housed in IntelliCages with access to water and chow *ad libitum*, and they were able to access bottles with a reward in the form of a saccharin solution (0.1% w/v), alcohol (4% w/v) or a mixture of the two. The probability of receiving a reward in two of the cage corners changed between 0.9 and 0.3 every 48 h over a period of ~33 days. We observed that in most animals, the odds of repeating the choice of a corner were increased if that choice was previously rewarded. Interestingly, in many cases, the time elapsed from the previous choice also increased the probability of repeating the choice, irrespective of the previous outcome. Behavioral data were fitted with a series of reinforcement learning models based on Q-learning. Then, extensions were introduced to account for attention or task complexity components or to directly include the interval length in order to simulate the decay of the expected value or to increase the probability of repeating the same choice. We find that introducing an interval-dependent adjustment to repeating the same choice allowed for the best prediction of the observed behavior, and the size of this effect may differ depending on the type of reward offered.

INTRODUCTION

Positive reinforcement increases the probability of repeating actions that were previously rewarded. An individual faced with a decision between two alternatives is more likely to choose the one with greater expected value. When the potential outcomes change due to shifts in the environment, it becomes necessary to sample the available choices and balance the exploitation and exploration of these choices. These behaviors are controlled by the brain's reward system, which signals prediction error and performs updates of expectation from any action or cue contingencies (Schultz, 2015). It has been observed that the strategy employed by humans, e.g., (Holroyd & Coles, 2002; O'Doherty *et al.*, 2003) and animals, e.g., (Fiorillo *et al.*, 2003; Bayer & Glimcher, 2005) when faced with a choice between probabilistic rewards is consistent with reinforcement models that rely on temporal difference learning and may be predicted with algorithms developed in the field of machine learning (Sutton & Barto, 1998). Reinforcement learning plays an essential role in adaptive behavior, and impaired decision making has been a major focus in research on the etiology of neuropsychiatric disorders (Maia & Frank, 2011).

Thus far, experimental models used in studies of reinforcement learning have been based on choices among actions with probabilistic outcomes made in environments with minimized distractions, short time scales and large numbers of choices performed in quick succession (e.g., (Clark *et al.*, 2004; Izquierdo *et al.*, 2017)). The effects of memory performance and attention on reinforcement learning have been taken into consideration in some published reports but mostly in the context of models where intervals between choices had limited influence (e.g., (Greggers & Menzel, 1993; Collins & Frank, 2012; Bai *et al.*, 2014; Collins *et al.*, 2017). Thus far, the effect of time on the action taken has usually been examined in the context of the interval between a salient cue and the associated outcome (e.g., (Fiorillo *et al.*, 2008; Gershman *et al.*, 2014)). Here, to observe reinforcement learning without strict task structure limitations, we tested animal behavior in a probabilistic choice reversal model in which mice were not compelled to perform the task in any way and choices were performed freely over a period of weeks. We tested two types of primary rewards, saccharin (0.1% w/v) and alcohol (4% w/v) solutions, which differ mechanistically in the way they affect the reward system. We found that most animals engaged in the task showed a moderate but significant preference towards higher reward probability. As anticipated, their choices were often significantly influenced by the outcome of the previous choice. However, we also observed that the interval between choices had an effect that was possibly independent of previous experience.

METHODS

ANIMALS

Experiments were performed on female C57BL/6J mice bred at the Institute of Pharmacology of the Polish Academy of Sciences in Krakow. Mice were housed in a conventional facility in Plexiglas cages (Type II L, 2–5 animals per cage) with aspen laboratory bedding (MIDI LTE E-002, Abedd) and nest building material. Breeding rooms had a 12 h light/dark cycle, with an ambient temperature of 22 ± 2°C and humidity of 40-60%. Animals were provided with a piece of aspen wood for chewing after weaning. Mice had *ad libitum* access to water and chow (RM1 A (P), Special Diets Services). All experiments were conducted in accordance with the European Union guidelines for the care and use of laboratory animals

(2010/63/EU). Experimental protocols were reviewed and approved by the II Local Bioethics Committee in Krakow (permits 1000/2012 and 1159/2015). Behavior was tested on female mice to reduce the risk of aggressive behaviors. The experimental groups are summarized in Table 1.

PROBABILISTIC CHOICE TASK

The IntelliCage apparatus (New Behavior, Switzerland) has a base made of transparent plastic (55 x 37.5 x 20.5 cm) with a metal cover and custom corner compartments. Each of the cage's corners is a small chamber that houses two 250 ml bottles, with nozzles accessible through guillotine doors. The size of the corner allows only one animal to enter the corner and access the bottles. Before being introduced to the IntelliCage, mice are implanted with radio frequency identification chips (RFID chips, UNO PICO ID, AnimaLab, Poland). An antenna inside the corner detects the chip and reports the animal number to the controlling software, which triggers preprogrammed events. The cage recorded the following parameters: temperature and luminosity in 1 minute intervals, presence of an animal in a corner (combined reading from a thermal sensor and an RFID antenna), crossing of photocell beams placed in the doors leading to the bottles, and lickometer contacts (the animal closing a circuit between the corner floor grating and the metal dipper of the bottle). Experiments were performed on groups of 14 female mice per cage. At the start of the test, mice were introduced to the IntelliCage, which had standard bedding and contained 4 plastic "houses" that the animals used as nests to sleep in during the day. The environmental conditions during the experiment were the same as those in the breeding rooms, with food and water available *ad libitum*. Exact schedules for each experiment are shown in Figure 1. At the start of the experiments, all corners had bottles filled with water. When a mouse entered a corner and the RFID chip was detected, both guillotine doors blocking access to the bottles would open with 0.5-s delay. The doors were closed when the mouse left the corner or 10 seconds after a lick of a bottle was detected. The initial period lasted between 4 and 7 days, and the mice were monitored daily to check whether all of them had learned to drink from the bottles. Then, the adaptation stage started, and a reward (saccharin 0.1% (w/v), alcohol 4% (w/v), a mixture of the two or plain water) became available in two of the corners. We chose a low concentration of alcohol to ensure high preference and to limit the effects of inebriation on learning. Saccharin was selected over sucrose to exclude nongustatory effects and to avoid clogging of the dippers. The adaptation stage lasted ~3 weeks, and bottle positions were changed regularly to prevent the formation of corner preferences (see Figure 1). Finally, during the main stage of the experiment, the probability of reward access varied between 90% and 30%, with a 2-s delay from the moment the animal was detected to the opening of the guillotine doors. Additionally, yellow LED lights in the reward corners were switched on when the animal was detected and switched off if the animal left the corner or after 2 s (irrespective of whether reward access was granted). The LED lights were intended as an additional cue of a choice being in progress to reduce the effect of variability caused by the animal detection mechanism. The positions of the corners with reward bottles were constant, while the probabilities changed between all possible states—90%:30%, 30%:90%, 90%:90% and 30%:30%—as shown in Figure 1.

DATA ANALYSIS AND STATISTICS

All analyses were performed using R (R Core Team, 2017). Basic differences between groups were assessed using the Kruskal-Wallis test followed *a posteriori* by the Dunn test adjusted for multiple comparisons using the Benjamini-Hochberg correction. One-sample comparisons were performed with

the Wilcoxon test. The significance (α) level was set at 5%. A complete set of raw data is provided as supplementary data.

Logistic regression was used to assess the effects of the outcome and time elapsed from the previous choice. The following formula was used:

$$\log\left(\frac{P(stay)}{(P(shift)}\right) = \beta_0 + \beta_1 \cdot interval + \beta_2 \cdot previous.outcome \qquad [1]$$

where β_0 is the intercept (preference independent of predictors), β_1 represents the effects of the time interval elapsed from the previous choice (per minute of interval), and β_2 is the effect of the outcome of the previous choice (binary, equals 1 after win). Regression was performed on a randomly selected ⅔ of the data points for an individual mouse, while the remaining ⅓ were used to test the goodness of fit of the model. The Wald test was used to assess the significance of the predictors.

Reinforcement learning modeling was performed using different modifications of the "basic" Q-learning (Watkins & Dayan, 1992):

$$Q_{a,n+1} = Q_{a,n} + \alpha(R - Q_{a,n}) \quad \text{[2] "basic"}$$

where $Q_{a,n}$ represents the expected value of choice a at step n, R is the reward (1 or 0) received after selecting choice a at step n, α is the learning rate parameter, and $Q_{a,n+1}$ is the expected value of action a at the next step. The first modification introduced was the "dual" learning rate, which depended on the R value:

$$Q_{a,n+1} = \begin{cases} Q_{a,n} + \alpha^{+}(R - Q_{a,n}) & if\ R = 1 \\ Q_{a,n} + \alpha^{-}(R - Q_{a,n}) & if\ R = 0 \end{cases} \qquad \text{[3] "dual"}$$

where α^{+} and α^{-} are separate learning rates for rewarded and nonrewarded choices. The second modification introduced an update for both the choice performed (a) and the non-selected choice (b) – a "fictitious" update (Hampton *et al.*, 2007):

$$Q_{a,n+1} = Q_{a,n} + \alpha(R - Q_{a,n}) \qquad Q_{b,n+1} = Q_{b,n} - \alpha(R - Q_{a,n}) \qquad \text{[4] "fictitious"}$$

A further extension of this approach is the "hybrid" model, which combines [3] and [4] (Cieślak *et al.*, 2018):

$$Q_{a,n+1} = \begin{cases} Q_{a,n} + \alpha^{+}(R - Q_{a,n})\ if\ R = 1 \\ Q_{a,n} + \alpha^{-}(R - Q_{a,n})\ if\ R = 0 \end{cases} \quad Q_{b,n+1} = \begin{cases} Q_{b,n} - \alpha^{+}(R - Q_{a,n})\ if\ R = 1 \\ Q_{b,n} - \alpha^{-}(R - Q_{a,n})\ if\ R = 0 \end{cases} \quad \text{[5] "hybrid"}$$

We also considered two models that add an additional step to equation [2]. The first model assumed that the difference between $Q_{a,n}$ and actual R may affect the learning rate; thus, the parameter η was

added to reflect an "attention" effect (Bai *et al.*, 2014). The models first adjusted the learning rate α_n based on prediction error and an η parameter (at η = 0, there is no attention effect):

$$\alpha_n = \eta\left|R - Q_{a,n}\right| + \alpha_n(1 - \eta) \qquad \text{[5] "attention"}$$

After the α_n update $Q_{a,n+1}$ was calculated using equation [2]. The second approach added a noise effect, ε, to the "forgetful" model, which depended on the number of alternatives available (Collins & Frank, 2012). In this case, equation [2] was followed by

$$Q_{a,n+1} = Q_{a,n+1} + \varepsilon(Q_0 - Q_{a,n+1})$$

$$Q_{b,n+1} = Q_{b,n+1} + \varepsilon(Q_0 - Q_{b,n+1}) \qquad \text{[6] "forgetful"}$$

where Q_0 is 0.5 (1 divided by the number of alternatives). As a way to directly introduce a memory "decay" effect to the reward, we came with the following equation:

$$Q_{a,n+1} = e^{-(t_{a,n} \cdot S)} \cdot \left(Q_{a,n} + \alpha\left(R - Q_{a,n}\right)\right) \qquad \text{[7] "decay"}$$

where n_a is the most recent step where choice a was made, $t_{a,n}$ is the respective time intervals (in minutes) between steps n_a or n_b and $n + 1$, and S is a parameter reflecting memory performance ($S = 0$ would indicate no memory decay). The memory decay component was based on observations in humans (Murre & Dros, 2015) and previous theoretical considerations (Woźniak *et al.*, 1995).

Models [2] to [6] were coupled with the softmax policy:

$$p(a,n) = \frac{e^{\beta \cdot Q_{a,n}}}{e^{\beta \cdot Q_{a,n}} + e^{\beta \cdot Q_{b,n}}} \qquad \text{[8]}$$

where $p(a,n)$ is the probability of choice *a* at step n, and β is an "inverse temperature" parameter that determines the extent to which the difference between expected values affects choice. In the case of the "decay" model, the policy included also a correction for the time effect on the expected value of the non-selected option (without updating the actual Q value), thus if a was selected at step $n - 1$ and $t_{b,n}$ is the time elapsed from the last step when b was selected, then:

$$p(a,n) = \frac{e^{\beta \cdot Q_{a,n}}}{e^{\beta \cdot Q_{a,n}} + e^{\beta \cdot (Q_{b,n} \cdot e^{-t_{b,n} \cdot S})}} \qquad \text{[9] "decay"}$$

where S is the same parameter as in equation [7].

Additionally, we propose a model in which the probability of repeating the same choice at step $n + 1$, $p(stay, n + 1)$, is modified by elapsed time:

$$p(stay, n+1) = \frac{e^{(b_0 + I \cdot t_{previous})}}{e^{(b_0 + I \cdot t_{previous})} + 1} \qquad [10] \text{ “time”}$$

where $t_{previous}$ is the time interval between $n+1$ and n, $b_0 = \frac{p(stay,n+1)}{1-p(stay,n+1)}$ at $t_{previous} = 0$, and I is a parameter controlling the effect of the interval length. Therefore, if the choice at step n was a, then $p(a, n+1) = p(stay, n+1)$. Finally, as a control, we considered a policy where $p(a,n) = p(b,n) = 0.5$ at every step, with or without interval-dependent adjustment (the “random” models).

Models were fitted using the Nelder and Mead simplex method (Nelder & Mead, 1965) implemented in the R *optim* function on a grid of starting points distributed as a Cartesian product for sequences from 0.05 to 1 by 0.05 for parameters: α, ε, η, S, and I, and from 0.25 to 15 by 0.05 for β. Parameters were limited to the range of (0, 1) for: α, ε, η, and S values and (0, 50) for β, and no limits on I. The optimal parameters for each model were selected from the optimization that scored the lowest sum of the negative logarithm of likelihood (nll):

$$nll = \sum_{n=1}^{N} -\log(p(chosen, n)) \qquad [11]$$

where $p(chosen, n)$ is the probability predicted by the model of the actually observed choice at step n. Model fits were compared using Akaike’s information criterion (AIC) (Akaike, 1974):

$$AIC = 2 \cdot nll + 2 \cdot k \qquad [12]$$

where k is the number of free parameters in the model.

RESULTS

AN UNCONSTRAINED PROBABILISTIC CHOICE TASK

We tested the behavior of group-housed female mice that could freely select between two probabilistic reward alternatives over an extended period of time (Figure 1). Three types of rewards were offered in separate experiments: a saccharin solution (0.1% w/v), alcohol (4% w/v), and a mixture of the two (alcohol+saccharin). As an additional control in a separate experiment, the reward was replaced with plain tap water. The experiments were split into two stages. The first was adaptation, during which the positions of the rewards were switched to reduce potential biases towards cage corners and to allow for the development of an alcohol preference. During the second, main stage, the positions of the bottles were fixed, and the probabilities of access to the bottles changed. As shown in Figure 2, the activity of animals, i.e., the total number of visits to the corners during each 48-h phase, was similar for all types of rewards and remained stable throughout the main phase of testing. Almost all animals showed a reward preference, which was calculated as the number of licks on the dippers of the bottles containing rewards, divided by the total number of licks on all bottles during the final 96 h of adaptation (Figure 3A). The median preferences were 84.6% for saccharin, 83.3% for the saccharin-alcohol mixture, and 62.9% for alcohol (all significantly higher than 50%, Figure 3A). No significant preference was

observed in the water control group, where the median was 39%. Note that 2 out of 13 mice showed no preference for alcohol+saccharin, and 5 of 28 lacked a preference for alcohol. Saccharin-containing solutions were preferred significantly more than alcohol (Figure 3A; Kruskal-Wallis test $H = 31.87$ $p = 5.55 \times 10^{-7}$). Further analyses were performed taking into account only visits that lasted sufficiently long for the outcome to occur (>2 s), which we designated as "choices". Over the course of the main stage experiment, the median number of choices was 1056 for alcohol, 1606 for the alcohol+saccharin mixture, 1252 for saccharin, and 1100 for the water control (Figure 3B). Considerable individual variation was observed, as the minimum total number of choices was 132 (alcohol) and the maximum was 2627 (alcohol+saccharin). The median number of choices was significantly higher in the alcohol+saccharin group than in the alcohol group (Figure 3B, Kruskal-Wallis test, $H = 9.192$, $p = 0.0268$). During the phases where mice chose between unequal probabilities, the preference for the alternative with a higher reward probability was 57.3% for saccharin and 58.9% for the alcohol+saccharin mixture, compared to 51.3% and 52.2% for alcohol and water, respectively (Figure 3C). In all these cases, the median preference values were significantly higher than random, and the preference for saccharin solutions was greater than the preference for alcohol or water (Kruskal-Wallis, $H = 33.808$, $p = 2.17 \times 10^{-7}$). The significantly higher-than-random preference for water suggests that the opening of the doors barring access to bottles became a conditioned reinforcer.

FACTORS AFFECTING CHOICE

We assessed whether the previous outcome correlated with the decision to repeat the choice ("stay") or select the alternative ("shift") by means of logistic regression with two predictors: the outcome of the previous choice ("win" or "lose") and the time interval from the previous choice. Examples of the regression curves are presented in Figure 4A&B. The first example (Figure 4A) corresponds to a mouse from a saccharin group, for which we note an effect of the interval on the probability of the "stay" behavior. As apparent from the distribution of the time intervals between choices, shorter, 1- to 5-minute intervals were more frequently observed in the case of "shift" decisions, while "stay' choices were more common at longer intervals, with a main peak at ~16 minutes (Figure 4A, upper panel). Accordingly, the logistic regression shows that the probability of a "stay" decision increased with the length of the interval, and the effect of previous outcome is noticeable as a shift of the curve (upward in the case of "win" vs. "lose", Figure 4A, lower panel). Conversely, in the second example drawn from the water control groups, there are no significant effects of interval or previous outcome (Figure 4B). A complete summary of logistic regression analyses for all mice is shown in Figure 5A-C. We found that in most cases, the models predicted a significant inherent propensity towards "stay" or "shift" responses, which was not accounted for by the predictors considered (the regression intercept, Figure 5A). Moreover, in most cases, the model showed significant effects of previous outcomes (Figure 5B) or interval length (Figure 5C), particularly in the alcohol+saccharin and saccharin groups. Curiously, in the alcohol group, there were four cases with opposite significant effects of interval, i.e., decreased probability of "stay" behavior at longer intervals. When only significant predictors were considered, the median values of the odds ratios were similar, and no significant effects of reward type were observed. The median value observed in the water control group compared to that of the other groups should be interpreted with caution, as it is based on very few cases of significant predictors in that group. A summary of the number of cases where the interval or previous outcome were found to be significant predictors is shown in Figure 5D. A significant effect of the interval was most frequent in the alcohol+saccharin group (10/13 cases), where it usually overlapped with the effect of previous outcomes (7/13). The overlap between a previous "win" and interval effects appeared to be less frequent in the alcohol and water control groups (3/28 and 0/14, respectively). To summarize, a

majority of mice were more likely to "stay" after a "win", and most animals were more likely to choose "stay" the when more time had elapsed from the previous choice. It should be noted, however, that the significant effect of the intercept may indicate that the predictors may not account for most of the variation observed.

REINFORCEMENT LEARNING

The preference for the choice associated with a higher reward probability and significant effects of previous outcomes as a predictor imply reinforcement learning, at least for some of the animals tested. Thus, we fitted a set of reinforcement learning models to find which model assumptions yielded the closest match to the observed behavior. All models considered were based on the Rescorla-Wagner equation, with the expected value of an action being updated with the observed prediction error (Rescorla & Wagner, 1972). The expected value was the weighted average of past outcomes, where recent events were weighted with a constant step-size parameter: the learning rate (the α parameter). Choice policy was based on the softmax function, which included an inverse temperature parameter (β). The simplest, "basic" model was equivalent to the Q-learning algorithm (Watkins & Dayan, 1992). "Dual" had separate learning rates for negative and positive values of the prediction error. The "fictitious" model simultaneously updated the value of both options (Hampton *et al.*, 2007), and "hybrid" added two learning rates to the fictive update (Cieślak *et al.*, 2018). The "attention" model adjusted the learning rate according to the prediction error, which may permit more optimal responses in a nonstationary environment (Bai *et al.*, 2014). Additionally, we considered models accounting for potential effects of memory performance or other effects of the time interval. First, the "forgetful" model introduced a component related to reverting to a base state of expected values (controlled by the "ε" parameter). The second model, "decay", was based on the assumption that the expected outcome values decay with time, following an inverse exponential function with the parameter *S* controlling steepness. The last two models were based on "basic" Q-learning, where the policy (rather than the expected value) was modified depending on the interval from the previous choice either using the time effect derived from the logistic regression fitting or directly finding an optimal value for the time effect parameter ("time+glm" and "time", respectively). Additionally, two reference models were considered: "random" choice, with fixed p = 0.5 for either alternative, and "time+random", where initial choice probabilities were fixed and then adjusted for interval length. The optimal parameter values were calculated for each model by finding the lowest sum of the negative log likelihoods (nll). The maps of the nll values from fitting the "basic" and "time" models to one animal from the saccharine-treated group are shown in Figure 6. The two sets of results are similar, though as apparent, the time-adjusted policy led to overall lower nll values. The results from the fitting of all the models were compared using AIC. The large variation in the AIC scores is mainly due to the differences in the number of choices per animal (Figure 3A). When median AIC values were compared, all models performed better than the "random" choice baseline (Figure 7). Surprisingly, it was the "time+random" model that had the lowest AIC scores for the largest number of animals (72 out of 83), while the "time" model scored best in remaining cases (11 out of 83). The median AIC difference was 3.8 between the 1st and 2nd best scoring models and 1108 between the 2nd and 3rd best scoring models. The result is particularly surprising considering the preference for the choice with higher reward probability (Figure 3C) and the logistic regression results (Figure 5B). We stress that while AIC scores (or other information criteria) are widely used as a method to compare model quality, the best fitting model is not necessarily the closest to the underlying mechanism employed by the reward system.

The optimal parameter values obtained for the "basic", "decay", and "time" models are shown in Figure 8. Reward-dependent significant differences in fitted α values were observed in the case of the "decay" model (Figure 8A, Kruskal-Wallis H = 9.1213, p = 0.02772), while the β values differed under the "basic" and "decay" models (Figure 8B: Kruskal-Wallis H = 16.384, p = 0.0009458 and H = 18.707, p = 0.0003143, respectively). However, the direction of the differences in the case of the β values were not consistent between models, and in the case of α under the "decay" model, a large number of values was at maximum (α=1), which prompts caution with the interpretation. The largest number of significant differences of fitted parameters were observed for the time effects (Figure 8C) The optimal median decay (S) was significantly lower in the water group compared to saccharin access.

Furthermore, there was a significant reward-dependent interval effect (I) in the "time" model, which was lower in the alcohol group (Figure 8C, Kruskal-Wallis H = 27.247, p = 5.225×10^{-6}). Optimal values of the interval effect (I) under "time+random" were almost identical as in case of "time", with the same difference in the median interval effect (I) in the alcohol group vs. all remaining groups (Figure 8C, Kruskal-Wallis H =26.952, p = 6.025×10^{-6}). Moreover, we note that the α and β values fitted in the "time" model were relatively low (median values smaller than 0.25), and in case of α, they had a similar range to the "basic" results. The fitted I values under "time" were much larger than those calculated from the logistic regression. In general, fitted values under the "time" model appeared to be the most consistent, with few cases of parameters at the maximum or minimum of the range.

DISCUSSION

We show reinforcement learning in mice under conditions where the animals are not compelled to perform the actions being reinforced. To our knowledge, this is the first study to demonstrate reinforcement learning under such circumstances. A critical difference between the paradigm presented here and previously described approaches is the lack of a trial and session trial structure usually employed in rodent or human studies. The number of attempts a mouse could make to receive a reward was unlimited, with only a short wait period (2 s) before the consequences of each choice were resolved.

As expected, when a choice was rewarded, the probability of repeating it was significantly increased in approximately half of the animals tested (33 out of 69, excluding the water control). The observed effect of previous outcomes appears to be weaker than that reported in studies employing classical models (e.g., (Frank *et al.*, 2004; Kwak *et al.*, 2014; Cieślak *et al.*, 2018)), though direct comparisons are limited due to major methodological differences. Furthermore, we found that the type of choice made correlated with the interval between decisions. First, as could be intuitively expected, a subsequent attempt was usually undertaken sooner when a choice was not rewarded than when it was rewarded. Second, the length of the interval correlated with an increased probability of repeating the same choice, irrespective of the previous outcome. This effect became apparent at intervals measured in minutes, and thus, it was probably negligible in previous studies where trials were repeated in quick succession. There was partial overlap between these two effects, and it remains unclear whether they are fully independent. A simple mechanistic explanation for the effect of the interval could be a decay of the memory of the expected value of choices. Since at the shortest intervals, the majority of choices were "shift", a decay of expected value could increase the probability of "stay". However, the simplistic assumption that expected outcome values decay to 0 over time is unlikely to account for the "stay" probability clearly exceeding 50% at longer intervals. Indeed, the comparison of reinforcement learning

model fits showed that the model incorporating a "decay" component was inferior in predicting choices to a model that adjusted the choice probability based on the interval. In fact, a model that included only the interval-dependent adjustment to the policy ("time+random") achieved dramatically better AIC scores than all models that did not. We fully acknowledge that the policy adjustment was arbitrary and may not necessarily reflect a part of the underlying mechanism. Nevertheless, the data clearly demonstrate an association between the type of choice made and the interval between choices. We note that our finding is limited to female C57BL/6J mice, and it remains unknown whether the effect may be generalized to other species. However, if generalization is possible, the finding could to an extent explain "irrational" behaviors, i.e., frequent choices of an inferior alternative.

A second objective of our study was to assess differences in the actions of saccharin and alcohol as reinforcers. The mechanisms by which alcohol and saccharin affect the reward system are inherently different. The effects should be instantaneous and transient in the case of saccharin (signaling from the gustatory system to the midbrain dopamine neurons (Simon *et al.*, 2006)) but are delayed by minutes and could be persistent in the case of alcohol (through activation of dopamine release and other mechanisms (Weiss *et al.*, 1993; Vengeliene *et al.*, 2008)). Nonetheless, it was reported that drug and natural rewards produce their long-term effects by acting on the same neuronal circuits (Kelley & Berridge, 2002; Pfarr *et al.*, 2018). Here, we found no evidence of impaired learning due to the effects of alcohol on memory performance; the behavior of mice in the saccharin and alcohol+saccharin groups was similar. In both cases, the rewards were highly preferred over water, and there was clear evidence that they induced reinforcement learning in most animals. Conversely, in the case of alcohol, while it was preferred over water after the 3-week adaptation stage, only some mice showed strong evidence of reinforcement learning. A possibility that should be considered is that the reward value of the alcohol solution was lower, which could be in line with the reported general preference of sweet taste over drugs in rodents (Ahmed, 2018). A low reward value would limit learning while possibly still being sufficient to produce a preference. We exclude the possibility that alcohol effects on memory consolidation could limit learning, as this was not the case for the alcohol and saccharin mixture. Additionally, there were at least some (5/28) animals that showed a significant increase in the probability of selecting a previously rewarded option. Furthermore, the experiment was not intended to model compulsive alcohol drinking, which persists despite increasing cost, reduced value or a risk of negative consequences (Vengeliene *et al.*, 2009; Hopf & Lesscher, 2014). Note that alcohol drinking under the experimental conditions presented here lacked the features of an alcohol addiction model. The concentration of alcohol was low; hence, blood ethanol levels (BEC) would be unlikely to match the values achieved in other published models (Rhodes *et al.*, 2005; Rodriguez Parkitna *et al.*, 2013). Nonetheless, if it is assumed that part of the mechanism involved in the development of addictions is a form of aberrant learning, then we note two observations. First, the "time" model that achieved the best fit to the observed behavior showed weaker effects of the interval for ethanol than for all other rewards (and the water control). A similar trend was apparent in the logistic regression analysis, where negative log odds ratio effects of the interval were obtained in some cases. Second, while in the case of almost all animals, optimal "time" models suggested very low inverse temperature parameters, there were outliers in the alcohol group. A considerably larger number of animals would have to be tested to conclude that alcohol produces a different form of learning in a subset of individuals, though we would argue that these results already provide some evidence for this idea. Moreover, our observation would be consistent with the observation that a subset of animals may prefer alcohol over sweetened water (Augier *et al.*, 2018). Finally, we note that previous reports linked alcohol preference to increased delay discounting (waiting impulsivity, (Sanchez-Roige *et al.*, 2014)). In the experiments presented here, there

was no correlation between alcohol preference at the end of the adaptation phase and the most probable interval length ($r = -0.063$ and $p = 0.5709$).

In conclusion, the most striking observation emerging from our results is that the effect of time elapsed from an action may affect the probability of repeating it to a considerable extent, and this effect may be dependent on the reward type.

ACKNOWLEDGEMENTS

This work was supported by the grant SONATA BIS 2012/07/E/NZ3/01785 from the National Science Centre, Poland and statutory funds of the Institute of Pharmacology of the Polish Academy of Sciences. JJ was supported with a PhD stipend from InterDokMed project no. POWR.03.02.00-00-I013/16. We would like to thank Dr. Rafał Bogacz for valuable comments and suggestions.

LITERATURE

Ahmed, S.H. (2018) Trying to make sense of rodents' drug choice behavior. *Prog. Neuropsychopharmacol. Biol. Psychiatry*, New directions in addiction, **87**, 3–10.

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. Autom. Control*, **19**, 716–723.

Augier, E., Barbier, E., Dulman, R.S., Licheri, V., Augier, G., Domi, E., Barchiesi, R., Farris, S., Nätt, D., Mayfield, R.D., Adermark, L., & Heilig, M. (2018) A molecular mechanism for choosing alcohol over an alternative reward. *Science*, **360**, 1321–1326.

Bai, Y., Katahira, K., & Ohira, H. (2014) Dual learning processes underlying human decision-making in reversal learning tasks: functional significance and evidence from the model fit to human behavior. *Front. Psychol.*, **5**.

Bayer, H.M. & Glimcher, P.W. (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, **47**, 129–141.

Cieślak, P.E., Ahn, W.-Y., Bogacz, R., & Rodriguez Parkitna, J. (2018) Selective Effects of the Loss of NMDA or mGluR5 Receptors in the Reward System on Adaptive Decision-Making. *eNeuro*, **5**.

Clark, L., Cools, R., & Robbins, T.W. (2004) The neuropsychology of ventral prefrontal cortex: Decision-making and reversal learning. *Brain Cogn.*, Development of Orbitofrontal Function, **55**, 41–53.

Collins, A.G.E., Ciullo, B., Frank, M.J., & Badre, D. (2017) Working Memory Load Strengthens Reward Prediction Errors. *J. Neurosci. Off. J. Soc. Neurosci.*, **37**, 4332–4342.

Collins, A.G.E. & Frank, M.J. (2012) How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *Eur. J. Neurosci.*, **35**, 1024–1035.

Fiorillo, C.D., Newsome, W.T., & Schultz, W. (2008) The temporal precision of reward prediction in dopamine neurons. *Nat. Neurosci.*, **11**, 966–973.

Fiorillo, C.D., Tobler, P.N., & Schultz, W. (2003) Discrete Coding of Reward Probability and Uncertainty by Dopamine Neurons. *Science*, **299**, 1898–1902.

Frank, M.J., Seeberger, L.C., & O'Reilly, R.C. (2004) By Carrot or by Stick: Cognitive Reinforcement Learning in Parkinsonism. *Science*, **306**, 1940–1943.

Gershman, S.J., Moustafa, A.A., & Ludvig, E.A. (2014) Time representation in reinforcement learning models of the basal ganglia. *Front. Comput. Neurosci.*, **7**.

Greggers, U. & Menzel, R. (1993) Memory dynamics and foraging strategies of honeybees. *Behav. Ecol. Sociobiol.*, **32**, 17–29.

Hampton, A.N., Adolphs, R., Tyszka, J.M., & O'Doherty, J.P. (2007) Contributions of the amygdala to reward expectancy and choice signals in human prefrontal cortex. *Neuron*, **55**, 545–555.

Holroyd, C.B. & Coles, M.G.H. (2002) The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychol. Rev.*, **109**, 679–709.

Hopf, F.W. & Lesscher, H.M.B. (2014) Rodent models for compulsive alcohol intake. *Alcohol*, Special Issue on Animal Models of Excessive Alcohol Consumption: Recent Advances and Future Challenges, **48**, 253–264.

Izquierdo, A., Brigman, J.L., Radke, A.K., Rudebeck, P.H., & Holmes, A. (2017) The neural basis of reversal learning: An updated perspective. *Neuroscience*, Cognitive Flexibility: Development, Disease, and Treatment, **345**, 12–26.

Kelley, A.E. & Berridge, K.C. (2002) The Neuroscience of Natural Rewards: Relevance to Addictive Drugs. *J. Neurosci.*, **22**, 3306–3311.

Kwak, S., Huh, N., Seo, J.-S., Lee, J.-E., Han, P.-L., & Jung, M.W. (2014) Role of dopamine D2 receptors in optimizing choice strategy in a dynamic and uncertain environment. *Front. Behav. Neurosci.*, **8**.

Maia, T.V. & Frank, M.J. (2011) From reinforcement learning models to psychiatric and neurological disorders. *Nat. Neurosci.*, **14**, 154.

Murre, J.M.J. & Dros, J. (2015) Replication and Analysis of Ebbinghaus' Forgetting Curve. *PloS One*, **10**, e0120644–e0120644.

Nelder, J.A. & Mead, R. (1965) A Simplex Method for Function Minimization. *Comput. J.*, **7**, 308–313.

O'Doherty, J.P., Dayan, P., Friston, K., Critchley, H., & Dolan, R.J. (2003) Temporal Difference Models and Reward-Related Learning in the Human Brain. *Neuron*, **38**, 329–337.

Rodriguez Parkitna, J.R., Sikora, M., Gołda, S., Gołembiowska, K., Bystrowska, B., Engblom, D., Bilbao, A., & Przewlocki, R. (2013) Novelty-seeking behaviors and the escalation of alcohol drinking after abstinence in mice are controlled by metabotropic glutamate receptor 5 on neurons expressing dopamine d1 receptors. *Biol. Psychiatry*, **73**, 263–270.

Pfarr, S., Schaaf, L., Reinert, J.K., Paul, E., Herrmannsdörfer, F., Roßmanith, M., Kuner, T., Hansson, A.C., Spanagel, R., Körber, C., & Sommer, W.H. (2018) Choice for Drug or Natural Reward Engages Largely Overlapping Neuronal Ensembles in the Infralimbic Prefrontal Cortex. *J. Neurosci.*, **38**, 3507–3519.

R Core Team (2017) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rescorla, R.A. & Wagner, A.R. (1972) A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Class. Cond. II Curr. Res. Theory*, **2**, 64–99.

Rhodes, J.S., Best, K., Belknap, J.K., Finn, D.A., & Crabbe, J.C. (2005) Evaluation of a simple model of ethanol drinking to intoxication in C57BL/6J mice. *Physiol. Behav.*, **84**, 53–63.

Sanchez-Roige, S., Baro, V., Trick, L., Peña-Oliver, Y., Stephens, D.N., & Duka, T. (2014) Exaggerated Waiting Impulsivity Associated with Human Binge Drinking, and High Alcohol Consumption in Mice. *Neuropsychopharmacology*, **39**, 2919–2927.

Schultz, W. (2015) Neuronal Reward and Decision Signals: From Theories to Data. *Physiol. Rev.*, **95**, 853–951.

Simon, S.A., de Araujo, I.E., Gutierrez, R., & Nicolelis, M.A.L. (2006) The neural mechanisms of gustation: a distributed processing code. *Nat. Rev. Neurosci.*, **7**, 890–901.

Sutton, R.S. & Barto, A.G. (1998) *Reinforcement Learning: An Introduction*, 1st Edition edition. edn. A Bradford Book, Cambridge, Mass.

Vengeliene, V., Bilbao, A., Molander, A., & Spanagel, R. (2008) Neuropharmacology of alcohol addiction. *Br. J. Pharmacol.*, **154**, 299–315.

Vengeliene, V., Celerier, E., Chaskiel, L., Penzo, F., & Spanagel, R. (2009) RESEARCH FOCUS ON COMPULSIVE BEHAVIOUR IN ANIMALS: Compulsive alcohol drinking in rodents. *Addict. Biol.*, **14**, 384–396.

Watkins, C.J.C.H. & Dayan, P. (1992) Q-learning. *Mach. Learn.*, **8**, 279–292.

Weiss, F., Lorang, M.T., Bloom, F.E., & Koob, G.F. (1993) Oral alcohol self-administration stimulates dopamine release in the rat nucleus accumbens: genetic and motivational determinants. *J. Pharmacol. Exp. Ther.*, **267**, 250–258.

Woźniak, P.A., Gorzelańczyk, E.J., & Murakowski, J.A. (1995) Two components of long-term memory. *Acta Neurobiol. Exp. (Warsz.)*, **55**, 301–305.

FIGURE LEGENDS

Figure 1. Experimental schedules. For each of the groups indicated on the left, the phases are represented with white, grey or black boxes. Black boxes indicate corners with full reward access (100%), dark grey boxes indicate a high probability of reward access (90%), light grey boxes indicate low value (30%) and white indicates access to water. The duration of each phase in hours is presented below the corresponding boxes.

Figure 2. Animal activity in the IntelliCages. The graphs show the total number visits in all corners of the cage per 48 h. Circles correspond to activity during the adaptation, and black points correspond to probability reversals. The ribbon shows the SEM values. Each graph summarizes the results from one group of animals (Table 1). The type of reward used in the experiment is indicated above.

Figure 3. Basic summary of choices. A. Reward preference during the last 96 h before the start of the reversals stage. Each dot corresponds to a single mouse with access to the substance indicated below the graph. B. Number of choices (visits longer than 2 s in corners with the reward) during the reversals stage. C. Fraction of choices corresponding to higher reward probability. The type of reward is indicated below the graphs. Each dot represents a single animal. Boxplots show medians, 1st & 3rd quartiles, and 95% confidence interval. Significant differences between the medians are indicated with stars ($p < 0.05$, Dunn's post hoc test with BH correction).

Figure 4. Examples of relations between choices and time intervals. A & B. Examples of behavior of a single animal from the saccharin and water (control) groups, respectively. The upper panels show the distribution of the two types of choices ("stay" or "shift") as a function of time elapsed from the previous choice. Each dot represents a single choice. Lower graphs show the logistic regression of the probability of "stay" depending on the previous outcome and interval.

Figure 5. Effects of previous outcome and time interval on choice. A-C. Graphs show the log odds ratio of "stay": independent of predictors, after a rewarded choice ("win"), and as a function of the interval, respectively. Each dot represents a single animal, with full dots marking a significant effect in the logistic regression. Boxplots show medians, 1st & 3rd quartiles, and 95% confidence interval, based only on significant results. The type of reward is indicated below the graphs. D. A summary of significant effects. The bars show the number of mice where predictors were found to be significant (Wald test). Note that the intercept from the regression is not included.

Figure 6. Example of negative log likelihood ("nll") value heatmaps from reinforcement model fitting. A & B. The maps show in grey the nll intensity as a function of α & β in the "basic" and "time" models fitted to the choices of a single mouse from the saccharin group.

Figure 7. Comparison of quality of reinforcement learning model fits based on AIC scores. A-D. The graphs show individual AIC values for all animals and all models fitted split based on the type of reward. Boxplots show medians, 1st & 3rd quartiles, and 95% confidence interval. The numbers above the model names indicate in how many cases the model achieved the lowest AIC value.

Figure 8. Optimal parameters from "basic", "decay", "time", and "time+random" models. Panel A&B show optimal values of α and β, respectively. C shows results for the time-effect parameters. Each dot represents a single animal, the bars show median values, and whiskers represent the 95% confidence interval. Models are labeled above the graphs, and reward type is indicated below the graph. Significant differences between medians are indicated with stars (*$p < 0.05$, ** $p < 0.01$, Dunn's post hoc test with BH correction).

TABLES

Table 1. Experimental groups.

Cohort	n	Reward	Starting age	Mean initial weight ± SEM [g]
II	14	Saccharin 0.1% (w/v)	8 weeks	19,49 ± 2,25
I	14	Saccharin 0.1% (w/v)	8 weeks	18,62 ± 1,79
II	14	Alcohol 4% (w/v)	8 weeks	19,04 ± 2,36
I	14	Alcohol 4% (w/v)	8 weeks	18,79 ± 1,64
III	14*	Saccharin+Alcohol	8 weeks	19,59 ± 1,95
IV	14	Water (control)	8 weeks	17,96 ± 0,19

*One animal lost its transponder and was excluded.

FIGURE 1

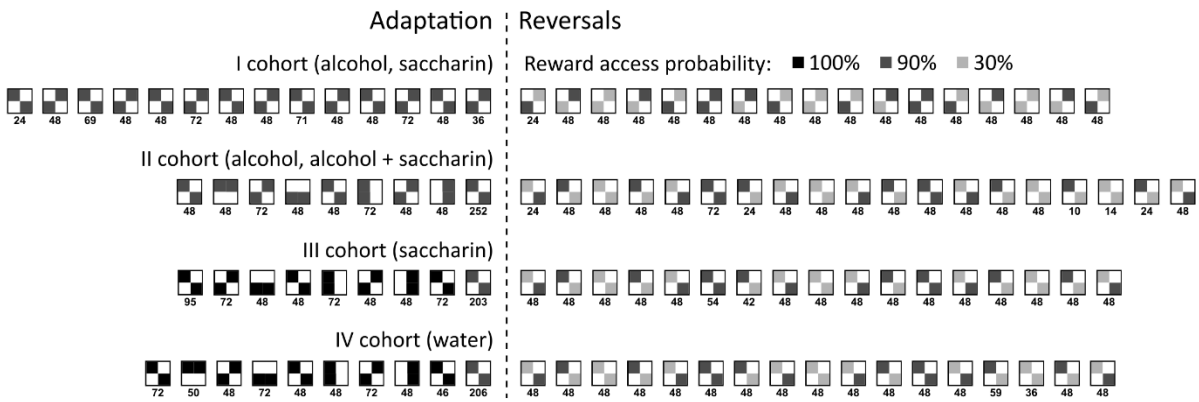

Figure 1. Experimental schedules. For each of the groups indicated on the left, the phases are represented with white, grey or black boxes. Black boxes indicate corners with full reward access (100%), dark grey boxes indicate a high probability of reward access (90%), light grey boxes indicate low value (30%) and white indicates access to water. The duration of each phase in hours is presented below the corresponding boxes.

FIGURE 2

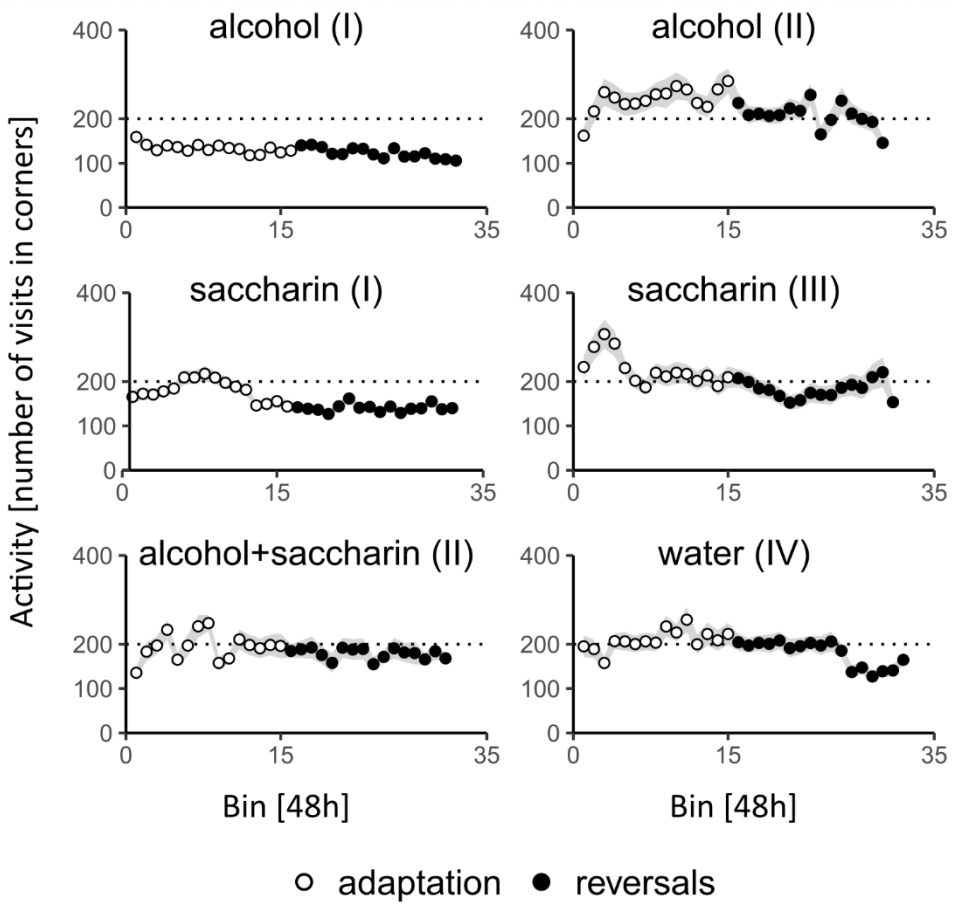

Figure 2. Animal activity in the IntelliCages. The graphs show the total number visits in all corners of the cage per 48 h. Circles correspond to activity during the adaptation, and black points correspond to probability reversals. The ribbon shows the SEM values. Each graph summarizes the results from one group of animals (Table 1). The type of reward used in the experiment is indicated above.

FIGURE 3

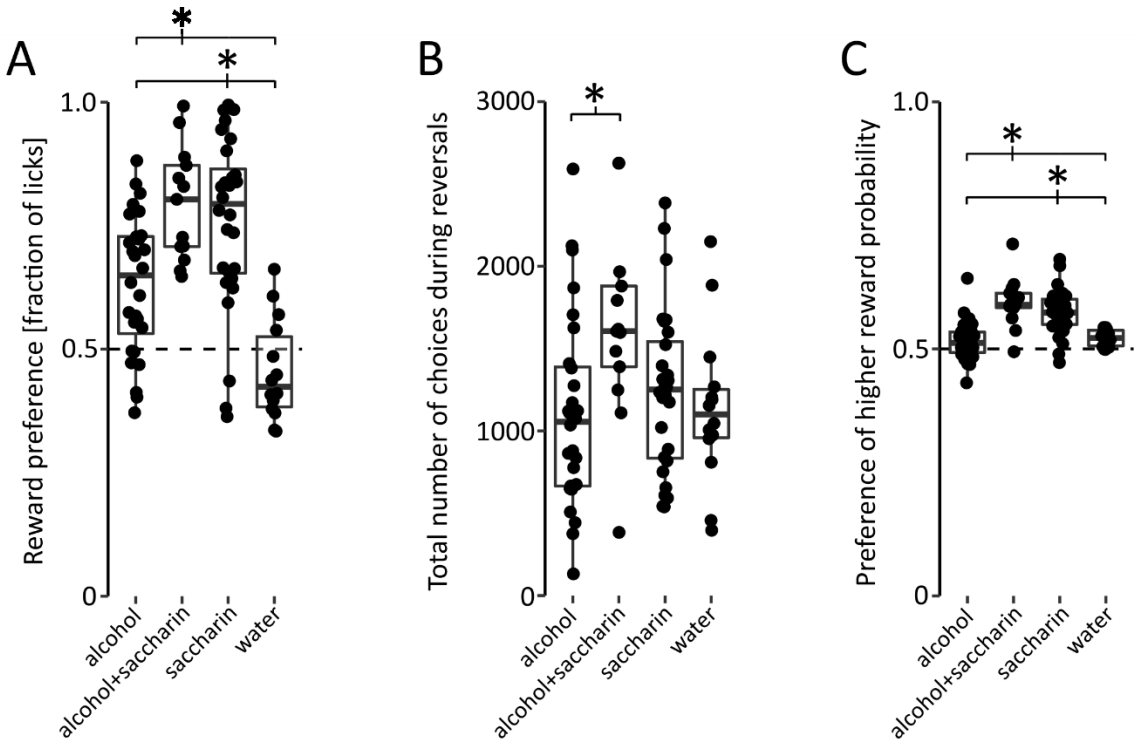

Figure 3. Basic summary of choices. A. Reward preference during the last 96 h before the start of the reversals stage. Each dot corresponds to a single mouse with access to the substance indicated below the graph. B. Number of choices (visits longer than 2 s in corners with the reward) during the reversals stage. C. Fraction of choices corresponding to higher reward probability. The type of reward is indicated below the graphs. Each dot represents a single animal. Boxplots show medians, 1st & 3rd quartiles, and 95% confidence interval. Significant differences between the medians are indicated with stars ($p < 0.05$, Dunn's post hoc test with BH correction).

FIGURE 4

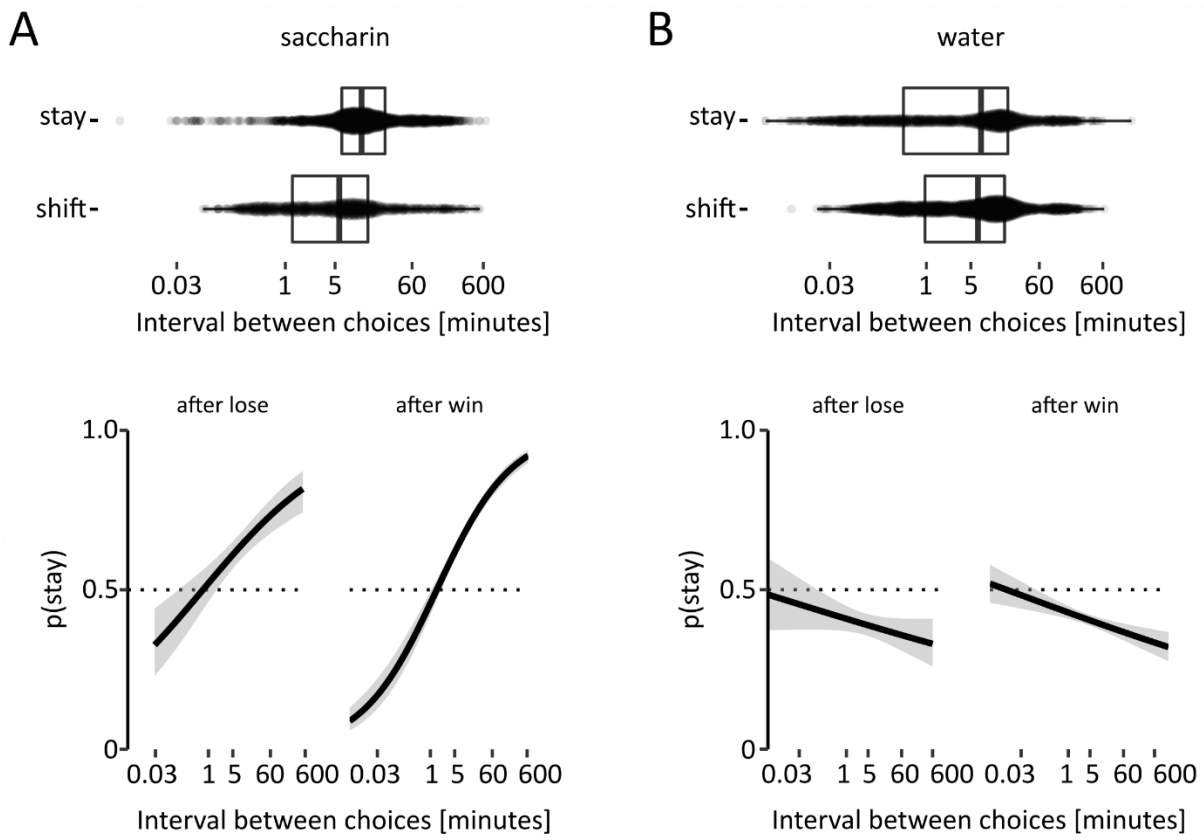

Figure 4. Examples of relations between choices and time intervals. A & B. Examples of behavior of a single animal from the saccharin and water (control) groups, respectively. The upper panels show the distribution of the two types of choices ("stay" or "shift") as a function of time elapsed from the previous choice. Each dot represents a single choice. Lower graphs show the logistic regression of the probability of "stay" depending on the previous outcome and interval.

FIGURE 5

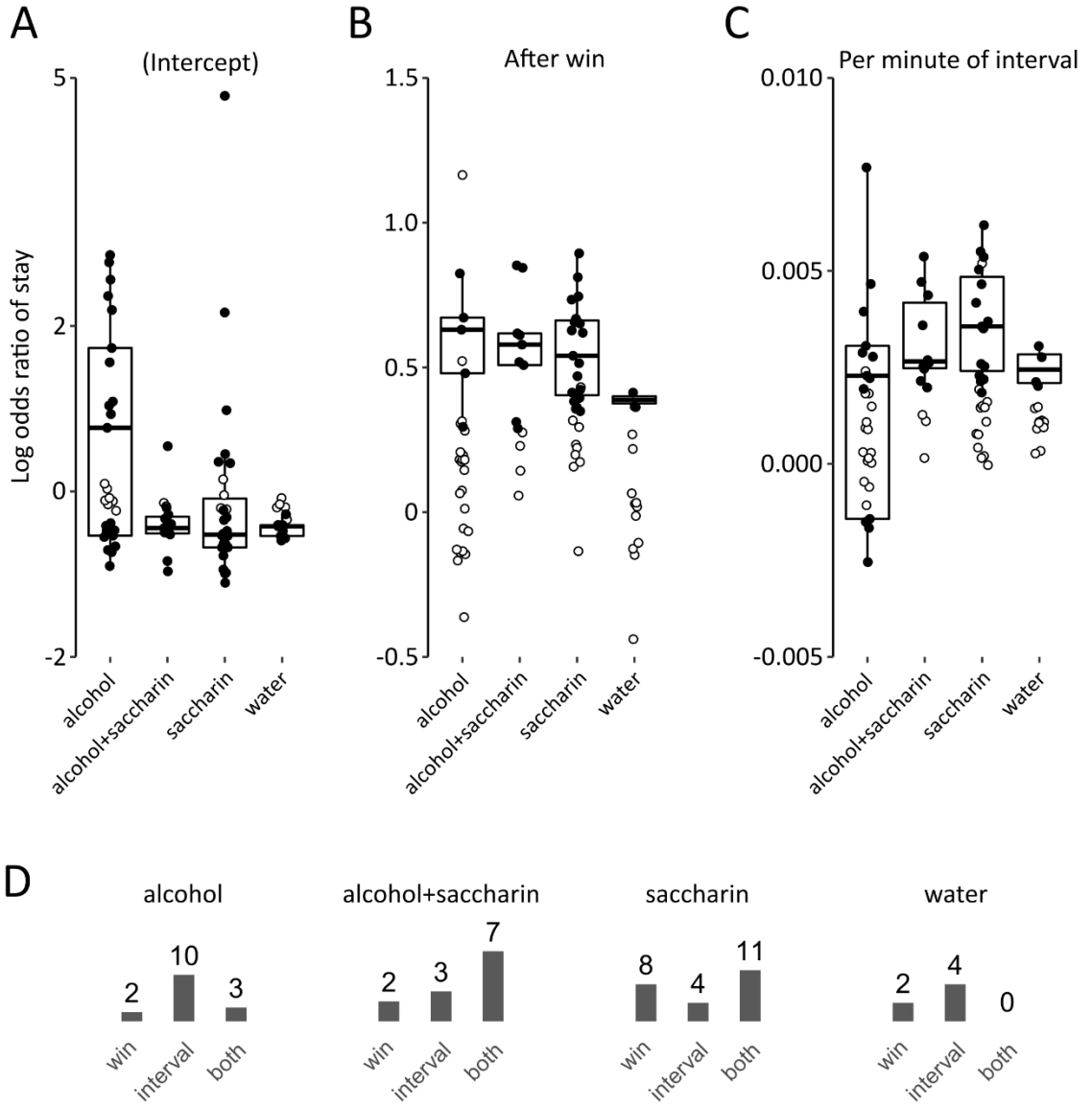

Figure 5. Effects of previous outcome and time interval on choice. A-C. Graphs show the log odds ratio of "stay": independent of predictors, after a rewarded choice ("win"), and as a function of the interval, respectively. Each dot represents a single animal, with full dots marking a significant effect in the logistic regression. Boxplots show medians, 1st & 3rd quartiles, and 95% confidence interval, based only on significant results. The type of reward is indicated below the graphs. D. A summary of significant effects. The bars show the number of mice where predictors were found to be significant (Wald test). Note that the intercept from the regression is not included.

FIGURE 6

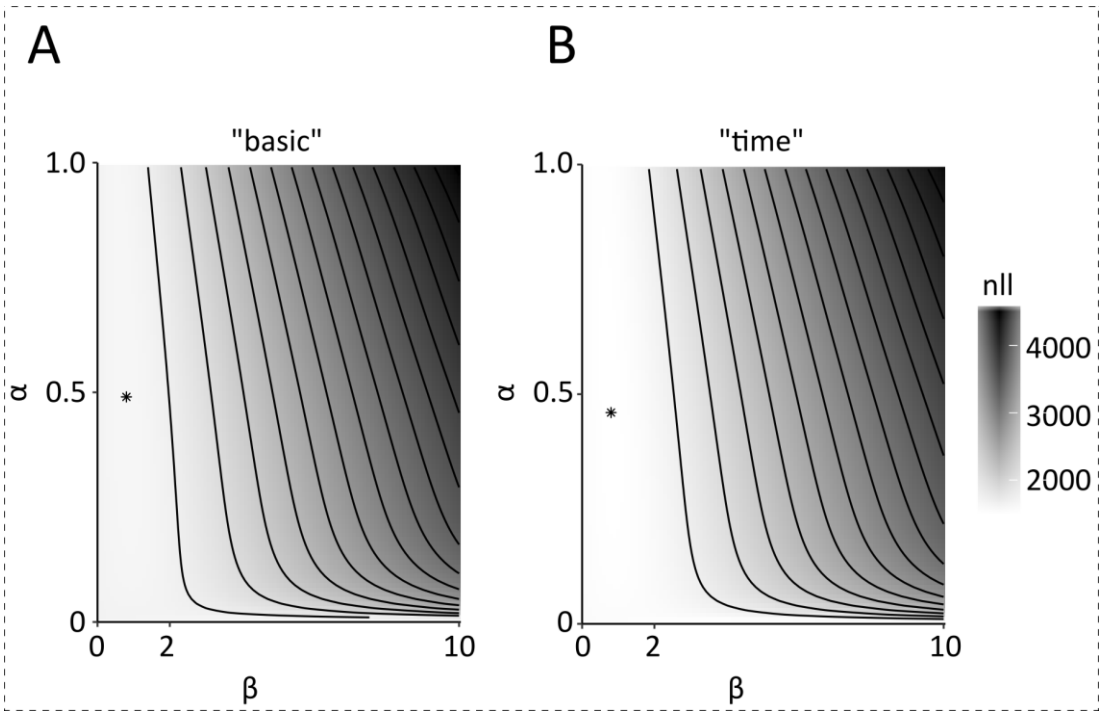

Figure 6. Example of negative log likelihood ("nll") value heatmaps from reinforcement model fitting. A & B. The maps show in grey the nll intensity as a function of α & β in the "basic" and "time" models fitted to the choices of a single mouse from the saccharin group.

FIGURE 7

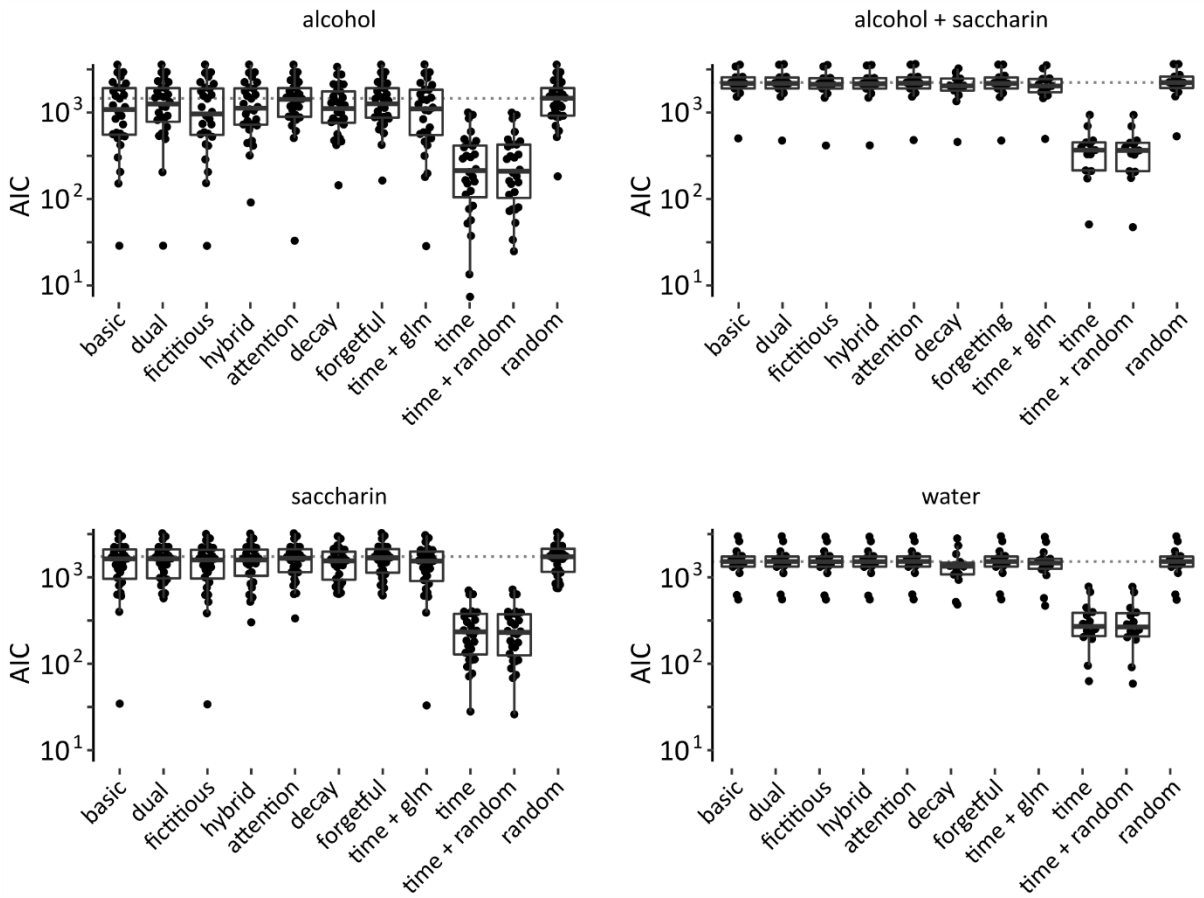

Figure 7. Comparison of quality of reinforcement learning model fits based on AIC scores. A-D. The graphs show individual AIC values for all animals and all models fitted split based on the type of reward. Boxplots show medians, 1st & 3rd quartiles, and 95% confidence interval. The numbers above the model names indicate in how many cases the model achieved the lowest AIC value.

FIGURE 8

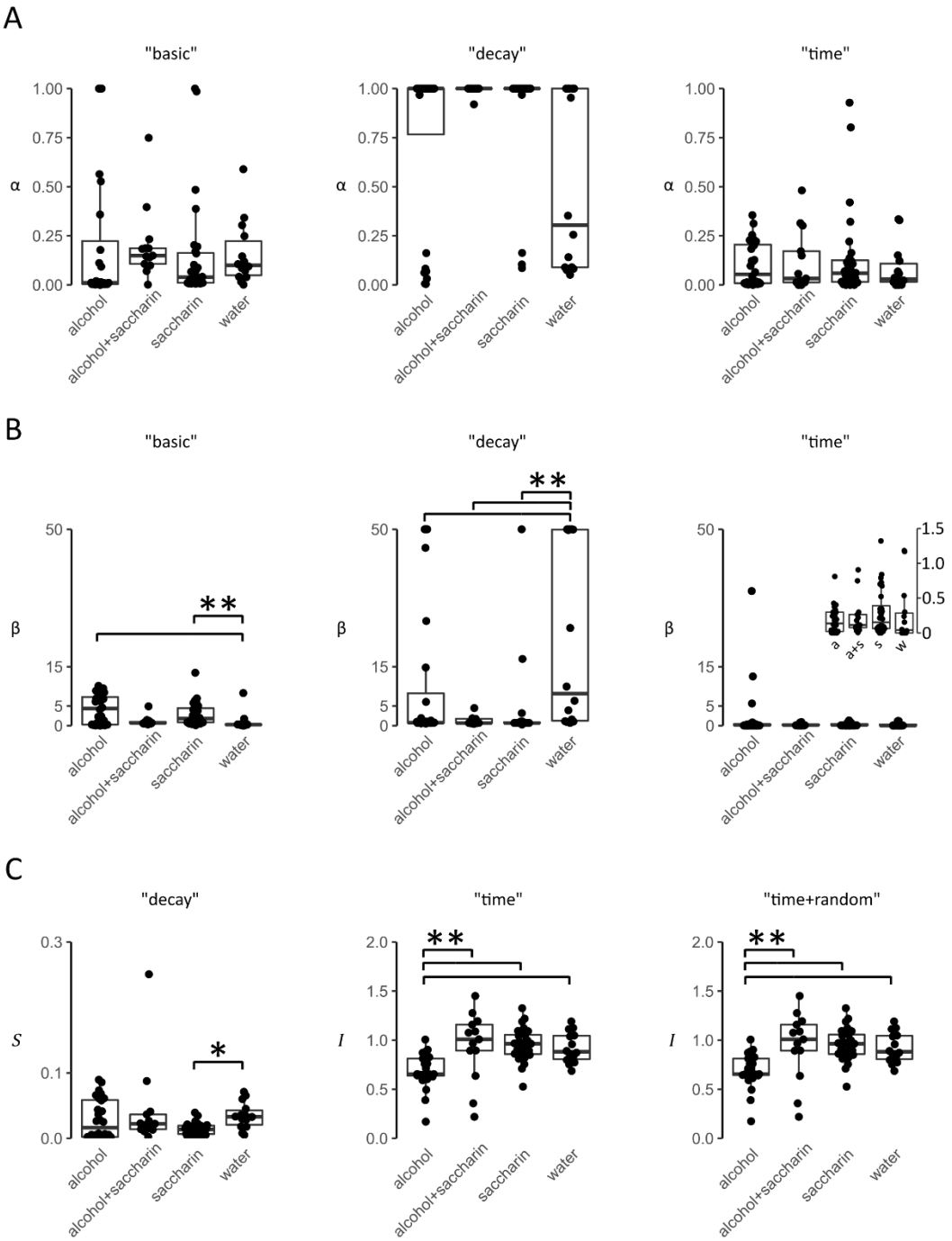

Figure 8. Optimal parameters from "basic", "decay", "time", and "time+random" models. Panel A&B show optimal values of α and β, respectively. C shows results for the time-effect parameters. Each dot represents a single animal, the bars show median values, and whiskers represent the 95% confidence interval. Models are labeled above the graphs, and reward type is indicated below the graph. Significant differences between medians are indicated with stars (*p < 0.05, ** p < 0.01, Dunn's post hoc test with BH correction).