The successor representation in high-risk drinking and alcohol-related contexts

M.P.M. Musial[1,2,3,4,*], S. Hall-McMaster[5], K. Shimomura[6,7], A. Kato[7,8], E. L. Bode[1,2,3], C. Grundmann[9], C. Ebrahimi[1], K. Morita[6,10], S.J. Gershman[5,11], T. Endrass[9], & F. Schlagenhauf[1,3,4]

[1]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Department of Psychiatry and Neurosciences | CCM, NeuroCure Clinical Research Center, Berlin, Germany

[2]Humboldt-Universität zu Berlin, Faculty of Life Sciences, Department of Psychology, Unter den Linden 6, 10099 Berlin, Germany

[3]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Einstein Center for Neurosciences Berlin, Berlin, Germany

[4]Charité – Universitätsmedizin Berlin, corporate member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Bernstein Center for Computational Neuroscience, Berlin, Germany

[5]Department of Psychology and Center for Brain Science, Harvard University, Cambridge, Massachusetts, United States of America

[6]Graduate School of Education, The University of Tokyo, Tokyo, Japan

[7]Japan Society for the Promotion of Science, Tokyo, Japan

[8]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York, United States of America

[9]Institute of Clinical Psychology and Psychotherapy, Technische Universität Dresden, Dresden, Germany

[10]International Research Center for Neurointelligence (WPI-IRCN), The University of Tokyo, Tokyo, Japan

[11]Center for Brains, Minds, and Machines, MIT, Cambridge, Massachusetts, United States of America

* Corresponding author

Abstract

The successor representation (SR) has been suggested to underlie nuanced forms of habitual behavior and a reduced SR variant (redSR) produces addiction-like behavior in simulations. Neither of these strategies can be detected in paradigms assessing habits in humans, which are usually conducted in disorder-irrelevant contexts, and this may explain inconsistent evidence for a goal-directed-to-habitual behavior shift in addiction. We tested whether individuals with high-risk drinking behavior rely more on (red)SR, particularly in alcohol-related contexts. Findings suggest that a (reduced) random-policy SR-like strategy contributes to human behavior, but that high-risk drinkers do not differ from low-risk drinkers in their use of this strategy. Instead, both groups rely less on (reduced) random-policy SR and more on model-free control in alcohol-related contexts. Results suggest that (reduced) random-policy SR supports adaptive, resource-efficient behavior and is selectively downregulated in substance-related contexts, highlighting the importance of contextual modulation in understanding decision strategies in mental health.

Introduction

When we unlock our phone's screen, we often do so in a quick and automatic way, not even planning to do anything specific on our device. This can be considered a habit [1] as it is inflexible, triggered by preceding stimuli, and does not depend on knowledge about action-outcome associations [1,2]. While habits save cognitive resources [3], they have been suggested to underlie repetitive behavior in mental health conditions, in particular substance use disorders [4–8]. The habit theory of addiction suggests that drug-related behavior is initially goal-directed but becomes more independent from its consequences and thus more habitual when transitioning to addiction [4,6,7].

Despite its intuitive appeal, evidence for this theory in humans is mixed [2,9–11], which might be due to how habits are usually measured. Whereas outcome devaluation and contingency degradation paradigms quantify habits as continued responding for an outcome which has been devalued or has become less contingent on the action [1], habitual and goal-directed behavior in the two-stage sequential decision-making task are often approximated by model-free (MF) and model-based (MB) reinforcement learning (RL), respectively [12–14] (but see [15]). Using these paradigms in addiction research comes with two limitations: they are conducted in disorder-irrelevant contexts, assuming that individuals with substance use disorders display habitual behavior beyond drug-related situations, and they conceptualize behavior as either goal-directed, habitual, or a mixture of both, neglecting third mechanisms [16].

One candidate strategy leading to more subtle inflexible behavior than that captured by MF RL or devaluation- or contingency-insensitive responding is the successor representation (SR) [17]. SR algorithms lie at an intermediate position between MB and MF RL in terms of behavioral flexibility and computational efficiency [18]. Unlike MF algorithms, which update the state value function under a given policy π directly, SR algorithms use a multi-step predictive matrix M^{π} representing each state by the discounted future occupancies of its successor states [18–20]. Specifically, each row of M^{π} represents a state s by how often we anticipate visiting any of its successor states s' under a given policy, discounting visits to successor states which took longer to reach using a time discount factor γ [20]. To illustrate, imagine a person who often drives their car home after work, but sometimes takes the subway to go for drinks. If they use SR RL, they represent their workplace by how often, starting from there, they anticipate visiting each successor state (car, home, subway, bar), no matter via which other states they got there. At decision time, they can then derive a value estimate for state s by a simple linear combination of the elements in row s of M^{π} and the rewards received per successor state s' contained in a

reward vector R (whose estimate *w* can be acquired through instruction or learned via prediction-error-based updates [21,22], see Methods). MB learners, by contrast, represent the environment via a one-step transition matrix T and need to iteratively combine knowledge about one-step transition probabilities (e.g., office-subway, subway-bar) with the respective elements in R to derive state value estimates.

SR's resource-efficient value estimation allows for flexible behavior under many circumstances. Like MB but unlike MF algorithms, SR RL can adapt to distal changes in reward (*reward revaluation*), also when these changes occur in a state which previously contained no reward (*goal-state revaluation*). Even when changes were not experienced starting from the current state (i.e., were distal), they can be reflected in *w* and linearly combined with the current state's row of M^{π}, resulting in updated value estimates. By contrast, SR use leads to inflexible behavior under certain conditions: Like MF but unlike MB algorithms, SR learners are unsuccessful in *policy revaluation*, meaning they cannot adapt to distal changes in rewards in states which have rarely been visited starting from the current state, e.g. as they contained no reward and the agent followed a near-optimal policy [20,21]. The new reward can be represented in *w* but will be multiplied by a discounted future occupancy near zero, leading to negligible effects on the current state's value estimate. Like MF but unlike MB algorithms, SR learners also cannot adapt to *transition revaluation*, i.e. distal changes in the transition structure [20,21], as transition experiences only affect the rows of M^{π} corresponding to their predecessor states. If altered transitions have not been experienced starting from the current state, they will not be reflected in the corresponding row of M^{π}. These inflexibilities are why SR RL has been described as a computational basis for a "subtler, more cognitive notion of habit" [21] than that produced by MF algorithms. However, since outcome devaluation and contingency degradation paradigms induce proximal changes in the reward structure and the two-step paradigm induces a type of reward revaluation, SR use leads to seemingly goal-directed behavior in these tasks (see [20]). Inconsistent evidence for increased habitual responding in addiction does thus not preclude that high-risk or addicted individuals rely more on SR strategies, a substrate for more subtle habitual behavior, than non-addicted individuals [21,23–26].

A theoretically appealing variant of SR RL is rigid, goal-based reduced SR (redSR) [22], where the predictive map M_{red}^{π} represents each state only by the expected future occupancies of salient states – typically goal-states such as those previously associated with drug reward (e.g., the bar) [22]. This dimension reduction conserves computational resources while maintaining sensitivity to rewards in goal-states, and may be adaptive under stable policies consistently leading to reward

[22]. However, this efficiency comes at the cost of behavioral flexibility: As M_{red}^{π} lacks columns representing non-goal-states' discounted future occupancies, and the weight vector w_{red} lacks elements for non-goal-states, redSR learners cannot reliably adapt to goal-state revaluation. In a simulation study, Shimomura et al. [22] examined how redSR agents behave when switching from a non-resistant policy (always pursuing drug rewards) to a resistant one (sometimes resisting drug-seeking). Even after repeated resistance, agents exhibited positive reward prediction errors (RPEs) upon reaching the goal state – paralleling drug-induced RPE signals hypothesized to reinforce drug taking [27].

Despite the habit-like characteristics of SR and redSR's phenotypic similarities to addiction [22], no study to date has investigated the extent to which individuals with high-risk drinking use SR or redSR and whether strategy use varies by context. One way to dissociate SR, redSR, MF, and MB strategies is by exploiting each algorithm's unique ability to adapt to reward, goal-state, policy, or transition revaluation [20–22]. When individuals experience all types of revaluation, we can test for signatures of SR use (higher success in reward and goal-state compared to transition and policy revaluation) and redSR use (higher success in reward than in goal-state revaluation) as opposed to MB and MF use (comparably high or low success across all revaluation types, respectively) [21]. To do so, we developed an active multi-stage decision-making paradigm based on a task introduced by Momennejad and colleagues [21]. In a preregistered (doi.org/10.17605/OSF.IO/9TUZE) cross-sectional online study, 520 alcohol users with high-risk (*n*=260) and low-risk (*n*=260) drinking behavior according to the Alcohol Use Disorder Identification Test (AUDIT) [28,29] performed our task in an alcohol- or non-alcohol-related context.

First, we aimed to replicate previous findings of SR use in humans [21] in a more lifelike task environment. Specifically, we hypothesized that 1) low-risk drinkers use a SR strategy to some extent when learning in a non-alcohol context, and that 2) their behavior cannot be explained by a MB strategy with different learning rates for rewards and transitions [21]. Based on the habit theory of addiction [4,6,7] and SR's inflexible properties [21], we further expected that 3) high-risk drinkers use SR to a larger degree than low-risk drinkers across contexts, but 4) particularly in an alcohol compared to a non-alcohol-related context. Lastly, given the resemblance of redSR-produced phenomena and addiction symptoms [22], we predicted that 5) in an alcohol context, high-risk drinkers rely more redSR compared to low-risk drinkers, and that 6) this group difference is more pronounced than in a non-alcohol context.

We found that humans use a form of SR, specifically a representation similar to a (reduced) random-policy SR [30]. Different than expected, this form of SR was not used more by high-risk

drinkers but was instead used less in alcohol-related contexts across low-risk and high-risk drinkers. The (reduced) random-policy SR seems to serve as an adaptive, resource-efficient representation which is used less in contexts that have previously been associated with drug use.

Results

SR, redSR, MB, and MF agents behave differently in a multi-stage decision-making task

Task design

Participants performed the active multi-stage sequential decision-making paradigm (Figure 1A, Figure S1) in either an alcohol context, i.e. a virtual bar with the goal to collect as many glasses of an individually preferred alcoholic drink as possible, or a non-alcohol context, i.e. a virtual apartment with the goal to collect as much hidden cash as possible (Figure 1B). The task consisted of 5 conditions. In each condition, participants had the chance to learn which out of four paths through a deterministic 10-state environment was optimal, i.e. led to the highest reward, in an initial learning phase. While all learning phase trials started from one initial state, trials in a subsequent re-learning phase started in later states. Here, participants experienced changes to the environment's reward structure (in reward, goal-state, policy revaluation, and control conditions) or transition structure (in the transition revaluation condition) which, in revaluation conditions, implied a change in optimal path departing from the environment's initial state. To assess how well participants adapted their path preference to the distal changes experienced during re-learning, we let them start from the initial state again with the instruction to maximize reward. The probability of changing path preference on such a test trial compared to the optimal path in the learning phase served as the primary measure of revaluation performance. A final rating phase, were participants were asked to indicate expected reward per initial-state-action, delivered a secondary, more deliberate measure of revaluation performance.

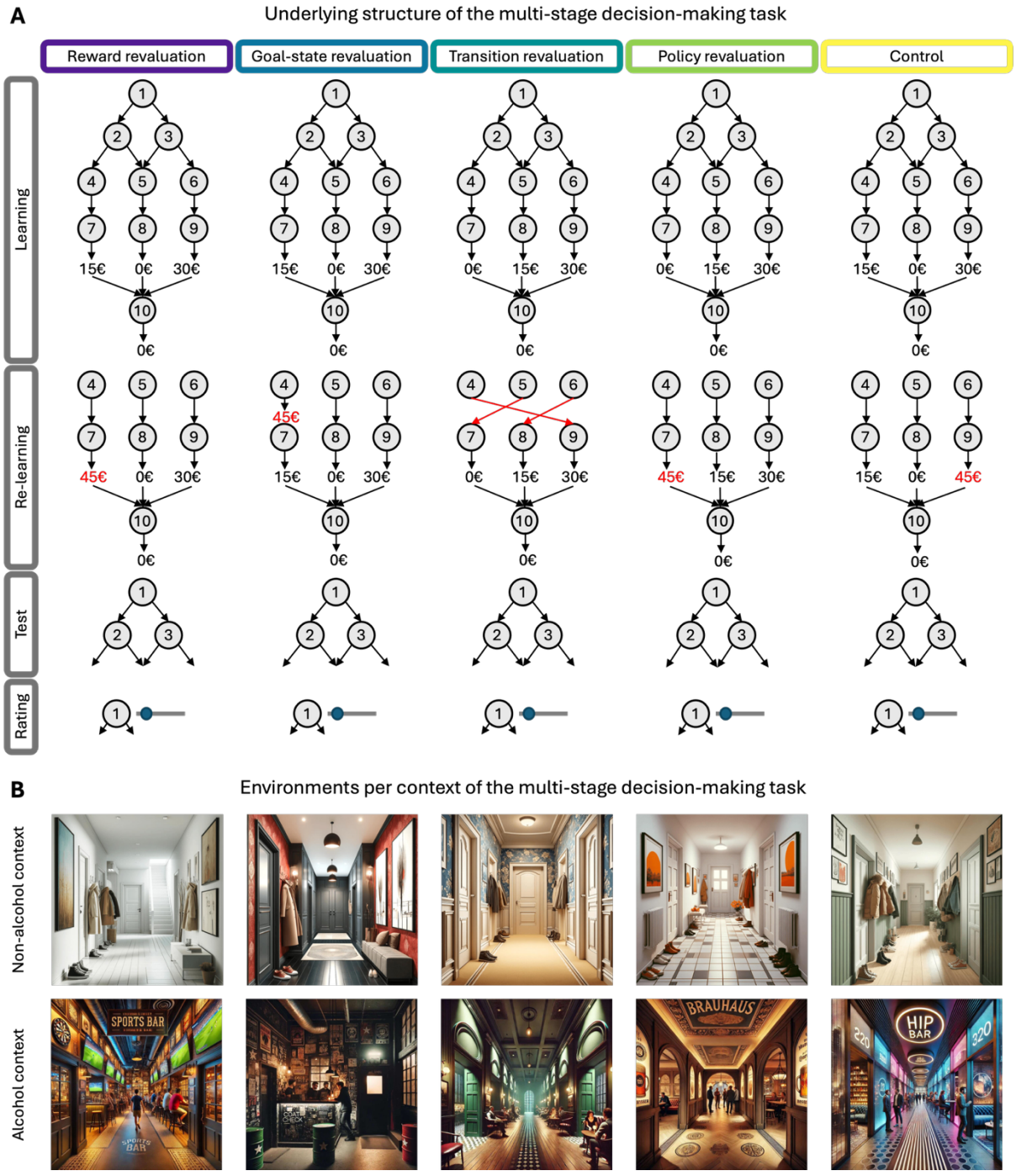

Figure 1. Multi-stage decision-making task. A Participants performed 5 task conditions (columns) encompassing 4 phases (rows) each. In each condition, they navigated through an environment consisting of 10 states (numbered circles) arranged in 5 stages and were instructed to maximize reward while staying vigilant to changes. In each state, participants had to perform an action (arrows) which deterministically led to a reward (€ amounts) and a subsequent state (except for state 10). A path through the environment was determined by the combination of state 1 and state 2 or 3 actions. During each phase, participants visited only the states shown. The initial learning phase encompassed 24 trials starting from state 1, out of which the first four were forced-choice, guaranteeing that each path was experienced once. Changes to the environment only occured in the subsequent re-learning phase, where all 9 trials started in third-stage states (3 trials from state 4, 5, and 6, each). In the reward revaluation condition, a new high reward was introduced following an action in a state which previously led to a small reward and should thus have been

visited with moderate frequency starting from state 1 during learning under a near-optimal policy. This condition implied a change in optimal state 1 action. In the goal-state revaluation condition, a new high reward was introduced following an action in a state which previously contained no reward but was positioned on the path to a small-reward-containing state and should thus also have been visited with moderate frequency starting from state 1 during learning under a near-optimal policy. Like reward revaluation, this condition implied a change in optimal state 1 action. In the transition revaluation condition, participants experienced a change in transitions between all third- and fourth-stage states. This condition implied a change in optimal state 1 action due to a change in the optimal second-stage action. In the policy revaluation condition, a new high reward was introduced in a state which previously contained no reward, was not positioned on a path to a reward-containing state and should thus have rarely been visited starting from state 1 during learning under a near-optimal policy. Like transition revaluation, this condition also implied a change in optimal state 1 action due to a change in the optimal second-stage action, but due to changes in the reward instead of the transition structure. The new reward introduced in the control condition did not imply a change in optimal path. Revaluation performance was assessed based on the subsequent 5-trial test phase, where participants were instructed to maximize reward starting from state 1. Test trials ended after the second-stage decision, meaning participants received no feedback on whether they had chosen the optimal path. In a final rating phase delivering a secondary measure of revaluation performance, participants were sequentially presented once with both state-1-actions and asked to indicate how much reward each action would lead to. Tasks in the alcohol and non-alcohol context followed the same depicted structure with the exception that, in the alcohol context, monetary rewards shown here were replaced by glasses of an individually preferred alcoholic drink (beer, wine, long drinks/cocktails) as follows: 0€ = 0 glasses, 15€ = 1 glass, 30€ = 2 glasses, and 45€ = 3 glasses. **B** Participants performed the multi-stage decision-making task in either an alcohol or a non-alcohol context. In each context, each condition took place in a different environment. Stimuli were generated using OpenAI's DALL-E. Here, only state 1 stimuli are shown. For full stimulus material, see publicly available code. For exemplary sequences from a trial, see Figure S1.

Simulations

To test whether SR, redSR, MB and MF RL produce the hypothesized behavioral signatures in the multi-stage decision-making task (Figure 2A), we simulated each algorithm's trial-by-trial behavior across conditions. Simulations confirmed our predictions under certain parameter settings (Figure 2B) but revealed nuances when using a range of other plausible parameter values for the learning rate α and the time discount factor γ (Figure S2). As expected, neither MF nor MB agents produced the hypothesized signatures of SR RL, i.e. higher reward and goal-state revaluation performance compared to transition and policy revaluation performance (near zero), at any considered parameter setting. However, SR itself produced these signatures only at γ=0.5 and redSR showed the same signatures at α=[0.5, 0.7] and γ=0.5 (for an explanation, see Methods). While redSR still exhibited its hypothesized signature, i.e. higher performance in reward compared to goal-state revaluation, at α=0.9 and γ=[0.5, 0.7], it could thus achieve flexibility comparable to SR RL despite its dimension-reduced form under other parameter settings. Note that this is also the case when assuming that redSR treats states beyond the previous reward-containing states as salient (Figure S3). Taken together, in the multi-stage decision-making task, the hypothesized signatures of SR use can be indicative of both SR or redSR while hypothesized marker of redSR use is specific to redSR.

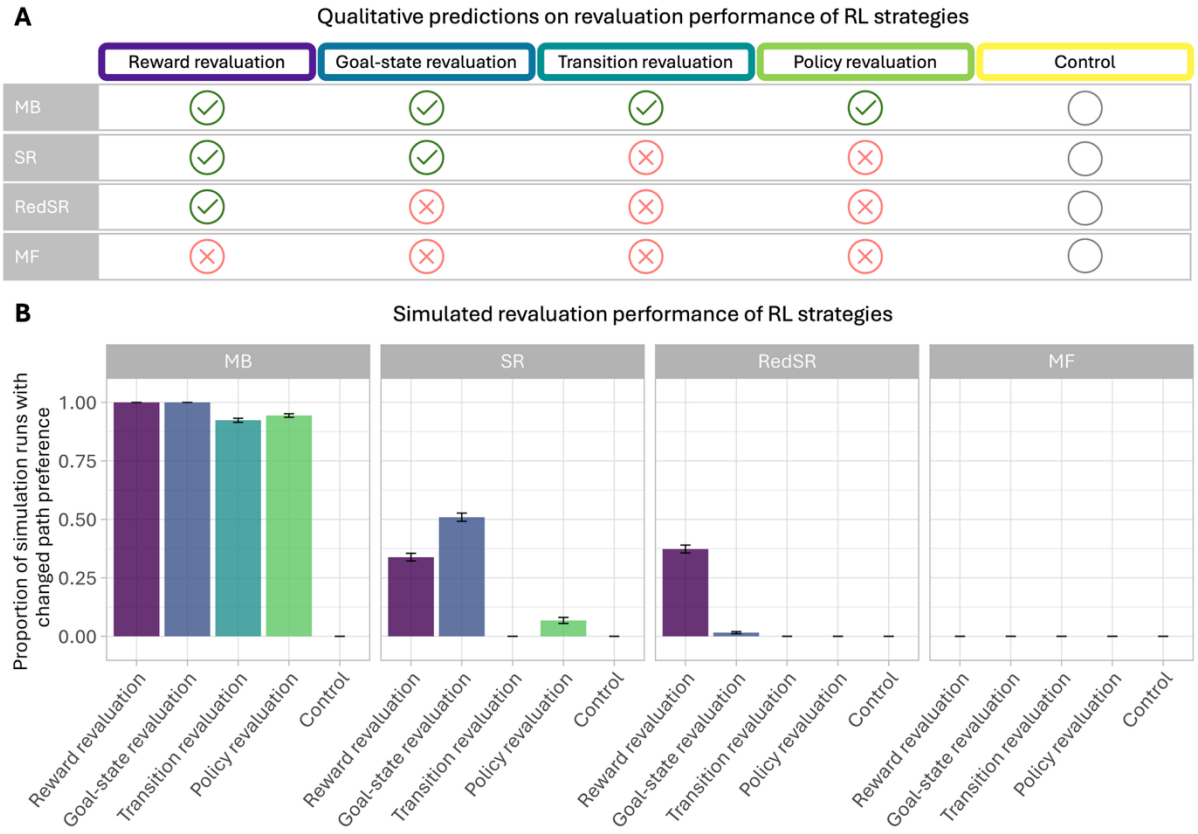

Figure 2. Predicted revaluation performance of *a priori* defined RL strategies.
A Qualitative predictions on whether model-based (MB), successor representation (SR), rigid goal-based reduced successor representation (redSR) with two goal-states, and model-free (MF) reinforcement learning (RL) algorithms (rows) can adapt to the different types of changes experienced in the conditions (columns) of the multi-stage decision-making task. Green ticks indicate that an algorithm is predicted to change its path preference in the test phase compared to the optimal path during the learning phase, whereas red crosses indicate it is not. Grey circles indicate that a change in optimal path from learning to test phase is not implied in the control condition. As shown, all RL strategies are expected to come with specific behavioral signatures: MB or MF RL show comparably high or low performance across revaluation conditions, respectively. One expected signature of SR agents is a lower revaluation performance in transition compared to reward and goal-state revaluation. A lower performance in policy compared to reward and goal-state revaluation is an additional predicted marker of SR RL and has been suggested to exclude that a higher revaluation performance in reward or goal-state compared to transition revaluation can be explained by a MB learner with a higher learning rate for the reward compared to the transition structure [21]. Lastly, the expected signature of redSR RL which differentiates it from SR RL is a lower revaluation performance in goal-state compared to reward revaluation. **B** Simulated revaluation performance of *a priori* defined reinforcement learning (RL) algorithms across task conditions, with 1000 runs per model and condition at α=0.9 and γ=0.5. Agents were considered to have an optimal path preference after learning if they had acquired a higher value for the optimal than the suboptimal state-1- and second-stage choice, respectively, and if they had chosen the optimal path in at least 3 out of the last 5 learning trials. Runs which fulfilled these criteria were taken as the base population to evaluate revaluation performance, i.e. the proportion of runs in which an agent changed its path preference on a test trial after re-learning compared to the optimal path during the learning phase. While, at these parameter settings, we found the expected model signatures, simulations at a range of parameters revealed more nuanced results (Figure S2). Bars represent means, error bars represent standard errors of the mean.

Low-risk drinkers use (reduced) random-policy SR in a non-alcohol context

Testing SR, redSR, MB, and MF contributions

After simulating the behavior of *a priori* defined RL strategies, we investigated their signatures in *n*=130 low-risk drinkers performing the task in a non-alcohol context (age=29.82±7.23 years; male/female/other: 60/70/0; AUDIT score=3.54±1.65; Table 1). Revaluation performance on a given test trial was predicted from task condition using logistic mixed-effects regression (for descriptive statistics, see Figure 3A, Table S1).

Low-risk drinkers were significantly more likely to be successful in reward (Hypothesis 1: *OR*=1191.93, *z*=6.24, *p*<.001) and goal-state revaluation (Hypothesis 1: *OR*=713.49, *z*=5.75, *p*<.001) compared to transition revaluation, respectively (Table S2). Considering that transition revaluation performance was not significantly higher than the probability of erroneous path preference changes in the control condition (*OR*=1.30, *z*=0.18, *p*=.855; Table S3), this pattern cannot solely be produced by MB, MF, or hybrid MB-MF use. While we did not observe the hypothesized redSR marker, i.e. lower performance in goal-state compared to reward revaluation (*OR*=0.60, *z*=-1.19, *p*=.236; Table S4), higher success in reward and goal-state compared to transition revaluation is a signature of SR and possibly also redSR use. Successful revaluation could not be explained by forgetting or exploratory behavior as performance in both reward (*OR*=1548.27, *z*=7.19, *p*<.001) and goal-state revaluation (*OR*=926.79, *z*=6.69, *p*<.001) was significantly higher than the probability of changes in path preference in the control condition (Table S3).

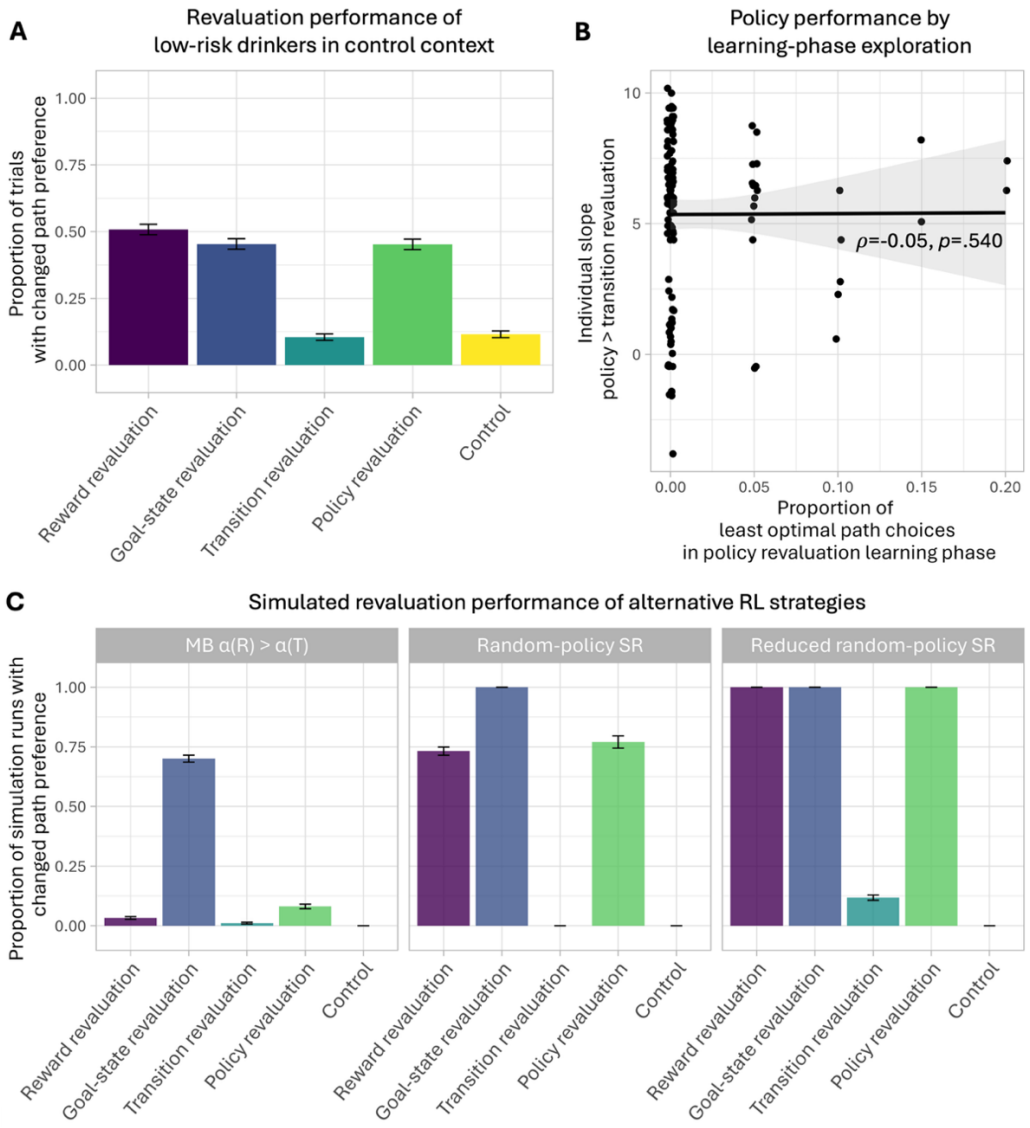

Figure 3. Revaluation performance in low-risk drinkers performing the multi-stage decision-making task in a non-alcohol context (*n*=130). A Observed proportion of test trials per condition in which participants changed their path preference, compared to the optimal path during learning. **B** Spearman's rank correlation between the proportion of learning trials in the policy revaluation condition on which participants chose the least optimal path (which would become optimal according to re-learning experiences) and their individual slopes for the contrast "policy > transition revaluation". Slopes were extracted from a logistic mixed-effects regression predicting the probability of changing path preference on a test trial, compared to the optimal path during learning, from condition. **C** Simulated revaluation performance of *post hoc* defined reinforcement learning (RL) algorithms across task conditions, with 1000 runs per model and condition at α=0.9 and γ=0.5. Agents were considered to have an optimal path preference after learning if they had acquired a higher value for the optimal than the suboptimal state 1 and second-stage choice and if they had chosen the optimal path in at least 3 out of the last 5 learning trials. Only runs which fulfilled these criteria were taken as the base population to evaluate revaluation performance, i.e. the proportion of runs in which an agent changed its path preference on a test trial compared to the optimal path during learning. The random-policy SR agent's higher performance in goal-state compared to reward and policy revaluation (like the SR agent's in Figure 2B) is due to a comparatively higher reward difference between the optimal and second-to-optimal path in goal-state revaluation. In **A** and **C**, bars represent means, error bars represent standard errors of the mean. In **B**, the line represents a linear fit and shaded grey areas represent the 95% confidence interval. MB, model-based; SR, successor representation; $\alpha(R)$, learning rate for the reward structure; $\alpha(T)$, learning rate for the transition structure.

One aspect of the observed pattern, however, was not in line with SR or redSR use: performance in policy revaluation was significantly higher than in transition revaluation (*OR*=765.00, *z*=6.00, *p*<.001; Table S2) and higher than the probability of erroneous path preference changes in the control condition (*OR*=993.70, *z*=6.92, *p*<.001; Table S3), but did not differ from performance in reward revaluation (Hypothesis 2: *OR*=1.56, *z*=1.31, *p*=.190; Table S5) or goal-state revaluation (Hypothesis 2: *OR*=0.93, *z*=-0.21, *p*=.833; Table S5) (Figure 3A). This pattern cannot be explained by redSR use as the factors underlying its unexpected success in simulations of goal-state revaluation do not apply when the newly introduced reward does not lie on a path to a previous goal-state (Methods). A standard SR strategy might produce high policy revaluation performance only if low-risk drinkers frequently chose the least optimal path during learning. We found no support for this explanation as the proportion of least optimal paths taken during learning was low ($M \pm SE$=0.01$\pm$0.003) and not correlated with individual slopes for the contrast between reward and transition revaluation performance (Figure 3B). It has previously been suggested that high reward and policy revaluation performance with concomitant low success in transition revaluation could be explained by a MB learner with a higher learning rate for the reward compared to the transition structure [21]. However, when simulating such a MB learner, we could not replicate the pattern observed in low-risk drinkers either (Figures 3C, S4).

Exploring alternative learning strategies

Given that none of the *a priori* defined models could explain the observed behavior, we explored alternative RL strategies. In fact, SR use allows for success in policy revaluation when making a simple modification: instead of learning M^{π} from experience, individuals could establish M^{random} under a uniform random policy [31]. This representation might be established during the initial forced-choice learning trials, which is supported by participants starting off with a high performance in the first free-choice trials (Figure S5). We simulated a random-policy SR agent which learned M^{random} and w in a MB way during forced-choice trials and did not update them until the announcement of the re-learning phase enabled updates to both structures again. This model, like low-risk drinkers, performed comparably well in reward and policy revaluation and had no success in transition revaluation across a range of parameter settings (Figure 3C, Figure S6). As this agent exhibited higher performance in goal-state compared to reward and policy revaluation, we explored whether a goal-based reduced version of random-policy SR would behave even more similarly to low-risk drinkers. Indeed, a reduced random-policy SR algorithm which established M^{random} by the end of the forced-choice trials, kept it constant throughout

the learning phase (whereas *w* was updated), and reduced it to a goal-based version assuming 4 goal-states (states 7-10) at the end of the learning phase produced equally high performance across reward, goal-state, and policy revaluation and low success in transition revaluation across a range of parameter settings (Figure 3C, Figure S7).

The (reduced) random-policy SR models are *post hoc* explanations which require assumptions e.g. about the rigidity of M^{random} during learning and the number of goal-states. While we do not claim that any of these exact strategies produced our data, some form of cognitive map which supports flexible adaptation to changes in policy and is not a one-step transition matrix T is necessary to explain low-risk drinkers' behavior (see Discussion). Independent of how a M^{random}-like representation was acquired, data suggest additional MF contributions which explain why performance in reward, goal-state and policy revaluation was below the level of simulated (reduced) random-policy SR models. A hybrid (reduced) random-policy SR-MF model is thus the most likely explanation of low-risk drinkers' behavior in a non-alcohol context.

High- and low-risk drinkers use less (reduced) random-policy SR in an alcohol context

Testing effects of high-risk drinking and substance-related context on SR, redSR, MB, and MF contributions

We tested effects of high-risk drinking and disorder-relevant context on strategy use based on additional data from low-risk drinkers performing the multi-stage decision-making task in an alcohol context (*n*=130; age=31.01$\pm$7.41 years; male/female/other: 67/63/0; AUDIT score=4.28$\pm$1.79) and from high-risk drinkers performing the task in either an alcohol (*n*=130; age=31.72$\pm$7.23 years; male/female/other: 84/46/0; AUDIT score=13.73$\pm$5.08) or non-alcohol context (*n*=130; age=30.94$\pm$7.48 years; male/female/other: 86/42/2; AUDIT score=15.34$\pm$5.27) (Table 1). Again, revaluation performance on a given test trial was predicted using logistic mixed-effects regression, this time from task condition, group, context, and their interactions (for descriptive statistics, see Figure 4, Table S1).

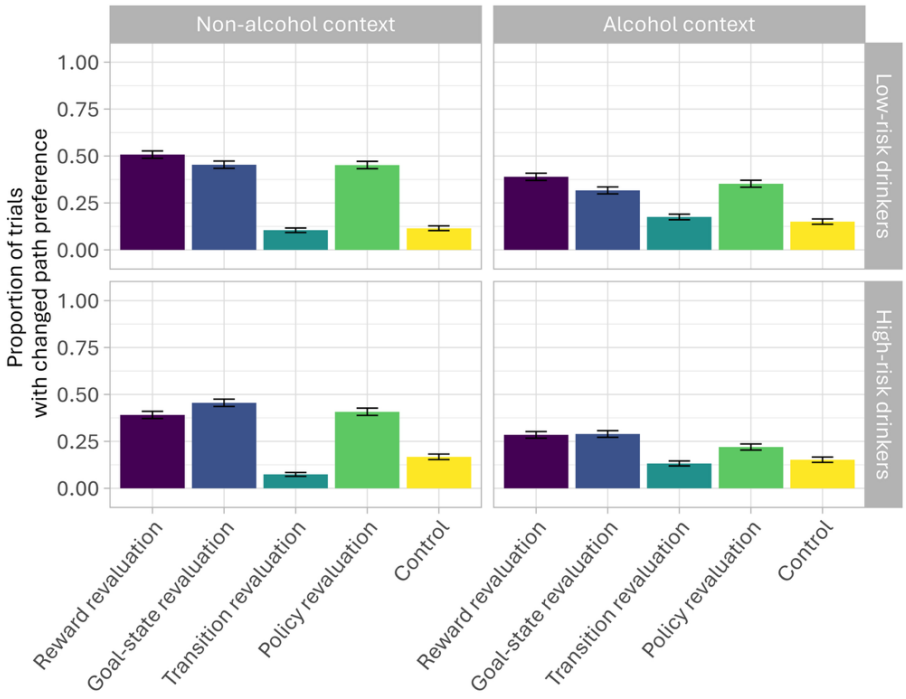

Figure 4. Revaluation performance in high-risk and low-risk drinkers performing the multi-stage decision-making task in an alcohol or non-alcohol context (*N*=520, *n*=130 per cell). Observed proportion of test trials per condition, group, and context on which participants changed their path preference, compared to the optimal path during learning. Bars represent means, error bars represent standard errors of the mean.

When testing for markers of SR or redSR use, we found that participants exhibited a higher revaluation performance in reward (*OR*=24.99, *z*=8.37, *p*<.001) and goal-state revaluation (*OR*=23.96, *z*=8.22, *p*<.001) compared to transition revaluation, respectively, averaged across groups and contexts (Table S6). Success in neither reward (*OR*=408.29, *z*=11.28, *p*<.001) nor goal-state revaluation (*OR*=391.40, *z*=11.34, *p*<.001) could be explained by random behavior or forgetting as it was more likely than erroneous path preference changes in the control condition averaged across groups and contexts (Table S7). While this was also true for performance in transition revaluation across contexts and groups (*OR*=16.34, *z*=4.54, *p*<.001; Table S7), the effect was comparatively small.

Contrary to our predictions, when testing for effects of group and context, we found that the contrast between reward and transition revaluation was not larger in high-risk drinkers compared to low-risk drinkers averaged over contexts (Hypothesis 3: *OR*=0.50, *z*=-1.56, *p*=.118), and the same was true for the contrast between goal-state and transition revaluation (*OR*=1.43, *z*=0.80, *p*=.426; Table S6). Instead, both contrasts were smaller, but still significant, in the alcohol context ('reward>transition': *OR*=8.55, *z*=4.95, *p*<.001; 'goal-state>transition': *OR*=7.02, *z*=4.45, *p*<.001) compared to the non-alcohol context ('reward>transition': *OR*=73.05, *z*=9.45, *p*<.001; 'goal-state>transition': *OR*=81.78, *z*=9.59, *p*<.001) when averaged across high-risk and

low-risk drinkers ('[reward>transition]*context': *OR*=0.12, *z*=-4.82, *p*<.001; '[goal-state>transition]*context': *OR*=0.09, *z*=-5.38, *p*<.001; Table S6). Context effects did not differ between groups (Hypothesis 4: '[reward>transition]*context*group': *OR*=1.18, *z*=0.19, *p*=.848; '[goal-state>transition]*context*group': *OR*=0.85, *z*=-0.18, *p*=.857; Table S6). While both high-risk and low-risk drinkers thus seemingly made use of a SR or redSR strategy, they did less so in an alcohol-related context.

Next, we tested for the signature of redSR use, i.e. lower performance in goal-state compared to reward revaluation. In the alcohol context, we did not find this marker averaged across groups (*OR*=0.94, *z*=-0.21, *p*=.834), and this did not differ between groups (Hypothesis 5: *OR*=2.58, *z*=1.90, *p*=.057; Table S8). Across both contexts and groups, we did not find a difference between goal-state and reward revaluation performance either (*OR*=0.96, *z*=-0.22, *p*=.822; Table S9). While an interaction with group (*OR*=2.87, *z*=2.87, *p*=.004) revealed that, across contexts, low-risk drinkers (*OR*=0.57, *z*=-2.23, *p*=.052) performed worse in goal-state compared to reward revaluation and thus made more use of redSR than high-risk drinkers (*OR*=1.62, *z*=1.80, *p*=.145), the contrast was not significant in any group (Table S9). We found no effect of context (*OR*=0.73, *z*=-0.83, *p*=.406) or of the interaction between group and context (Hypothesis 6: *OR*=0.72, *z*=-0.46, *p*=.647) on the contrast between goal-state and reward revaluation performance (Table S9). In conclusion, the data do not carry unique signatures of redSR use in any group or context.

Exploring effects of high-risk drinking and substance-related context on alternative learning strategies

While the observed higher performance in reward and goal-state compared to transition revaluation could indicate SR or redSR use, the full pattern across contexts and groups can only be explained by a strategy similar to (reduced) random-policy SR: participants were significantly more successful in policy compared to transition revaluation (*OR*=19.64, *z*=7.88, *p*<.001). Again, this was less pronounced in the alcohol (*OR*=5.72, *z*=4.12, *p*<.001) compared to the non-alcohol context (*OR*=67.47, *z*=9.55, *p*<.001; '[policy>transition]*context': *OR*=0.08, *z*=-5.86, *p*<.001; Table S5). We found no interactions between the contrast between policy and transition revaluation performance and group (*OR*=0.64, *z*=-1.08, *p*=.282) or group and context (*OR*=0.54, *z*=-0.74, *p*=.459; Table S5). At the same time, the probability of success in policy revaluation was not significantly different from that in reward revaluation across groups and contexts (*OR*=0.79, *z*=-1.28, *p*=.202), and this did not differ significantly depending on group (*OR*=1.28, *z*=0.67, *p*=.503), context (*OR*=0.72, *z*=-0.87, *p*=.386), or their interaction (*OR*=0.46, *z*=-1.09, *p*=.277; Table S9). Again, this pattern could not be explained by a standard SR strategy with high exploration

during learning, as we found no association between the proportion of policy revaluation learning trials where the least optimal path was chosen and individual slopes for the contrast between policy and transition revaluation in the full sample (ρ=.001, p=.989).

Taken together, results suggest that a (reduced) random-policy SR-like representation shapes behavior not only in low-risk drinkers performing the task in a non-alcohol context but also in high-risk drinkers and alcohol contexts. High-risk drinking is not associated with more (reduced) random-policy SR use, which is in line with exploratory results showing no significant associations between use of this strategy and self-report measures of e.g. drinking frequency or habitual alcohol intake (Figure S8). Instead, participants make less use of this strategy in substance-related contexts. Reward, goal-state, and transition revaluation performance lie below simulated performance of a (reduced) random-policy SR strategy, suggesting a hybrid (reduced) random-policy SR-MF model as the most likely explanation for behavior across groups and contexts, with larger MF contributions in alcohol compared to non-alcohol contexts. Importantly, this interpretation holds when measuring revaluation performance via a more deliberate judgement of value as we could replicate all effects from test phase analyses in rating phase data (Figure S9, Tables S10-S18). For a discussion of potential alternative explanations and confounding factors, see Supplement.

Discussion

The habit theory of addiction predicts that the transition from recreational drinking to substance use disorders is associated with enhanced habitual and reduced goal-directed behavior [4,6,7], but evidence in humans is inconsistent [2,9–11]. We suspected two reasons for this: First, high-risk drinkers might make increased use of SR [17,18] and redSR [22] RL, two strategies with subtle, habit-like characteristics [18,21] which are not detectable in common paradigms used to assess habitual behavior [1,12], and second, habit paradigms are conducted in disorder-irrelevant contexts. Here, we thus set out to investigate whether SR and redSR use is increased in high-risk drinkers making disorder-relevant choices.

In a first step, we aimed to replicate previous findings suggesting that humans use SR [21,23,25,32]. With an extended version of the multi-stage decision-making task by Momennejad et al. [21] in a more naturalistic non-alcohol-related context, we found that the behavior of 130 low-risk drinkers could not be explained by a MB, MF, or a hybrid MB-MF strategy alone: Participants exhibited higher reward than transition revaluation performance – a key signature of SR use [20,21] which, according to our simulations, could also indicate redSR use [22]. Unlike in [21], our sample

was no more likely to adapt to transition revaluation than to erroneously change path preference. This is not in line with a hybrid SR-MB strategy like SR-Dyna [20,21], where MB contributions should allow some success in transition revaluation, nor with a recently suggested probabilistic SR [33] or a default representation updated via matrix inversion identities [31]. In light of previous findings indicating that humans rely more on a MB strategy when SR leads to incorrect value estimation [26] and MB is thus most needed [21], the low MB engagement in transition revaluation seems surprising. However, participants did not experience that SR-derived value estimates for state-1 actions were inaccurate (unlike in [26]), and the larger state space compared to [21] enhanced the difference in computational costs between MB and SR RL and likely cognitive load [34]. A cost-benefit arbitration would thus have favored SR over MB use in our task.

While the key signature of SR or redSR use was more pronounced in our data compared to [21], neither of these policy-dependent strategies can explain why low-risk drinkers did not perform worse in policy than in reward revaluation in a non-alcohol-related context. Adaptation to distal new rewards introduced in a state which has rarely been visited in the past suggests some form of policy-independent representation. As we concomitantly saw low performance in transition revaluation, this cannot be the one-step transition matrix T used in MB RL. Instead, use of a random-policy SR-like map M^{random} which represents all paths through the environment with equal expected occupancies offers a plausible explanation [31]. The introduction of forced-choice trials, unlike in [20], enforced equal experience of all paths and might have favored the formation of M^{random}. While a single experience per path would likely not be enough for a TD-based SR strategy to learn M^{random} until convergence, participants might have used a computationally expensive MB strategy at the beginning and then transformed T to M^{random} to benefit from cheaper computations throughout the learning phase. Such a strategy could reproduce key characteristics of the observed revaluation pattern when simulated and would align with evidence that humans and rats use a more MB strategy in earlier task trials before shifting towards SR use [35]. A neurally plausible mechanism to learn M^{random} assuming no transformation between MB and SR representations might be based on Behavioral Timescale Synaptic Plasticity (BTSP) [36–38]. BTSP enables a representation of each state in terms of sequences in which it was encountered [36,39], and the dot product of two states' representations, similar to one element in M^{random}, quantifies how many sequences both states are part of [36]. As a sequence (here, the states experienced during one trial) can be learned in one shot, a representation like M^{random} can be formed during four trials covering the full environment [36–38]. This and other recent findings [40] suggest that neither MB nor TD learning are necessary to rapidly

establish an SR-like cognitive map which, if established during forced-choice trials, would look like M^{random} in our task.

We found that the (reduced) random-policy SR-like strategy can explain key characteristics of behavior also in high-risk drinkers and alcohol-related contexts. Different than expected, SR (or redSR) use was not more pronounced in high-risk compared to low-risk drinkers across contexts, nor was a unique marker of redSR use more pronounced in high-risk drinkers when exclusively considering the alcohol context. This suggests that the inconsistent evidence for increased habitual responding in human addiction [2,9,10] is unlikely due to a shift from MB to SR-based strategies which cannot be detected in classic habit paradigms [1,12]. Instead, while (reduced) random-policy SR use leads to habit-like inflexibility when distal transitions in the environment change, it is generally an adaptive and flexible strategy. Random-policy SR-like models thus do not seem to be a suitable formalization of dysfunctional, inflexible behavior and, from the perspective of the habit theory of addiction [4,6,7], it is thus unsurprising that individuals with more risky and automated drinking behavior (Table 1) do not make more use of such strategies. Simulations showed that even goal-based dimension reduction, when applied to a random-policy SR, allows for flexible, policy-independent adaptation to changes in the environment's reward structure and captures key characteristics of the observed revaluation pattern. To further investigate the role of dimension reduction in high-risk drinking, future studies should use a type of goal-state revaluation where new rewards appear after previous goal-states (Methods) and explore the effect of resistance to goal-reaching, either by experimentally inducing it or focusing on individuals wanting to limit alcohol consumption [22].

Different than expected, individuals relied less on a (reduced) random-policy SR-like strategy and more on MF learning in an alcohol-related context. Alternative explanations for context effects can largely be ruled out, as we ensured comparable learning in both contexts, participants did not behave more randomly in the alcohol-related context, and each context included five different environments, minimizing effects of specific stimuli. Our finding suggests that, in individuals with prior alcohol experience, (reduced) random-policy SR-like strategies are selectively downregulated in alcohol-related contexts. It has previously been suggested that habitual and goal-directed contributions to behavior depend on each strategy's certainty in a context-dependent manner [41]. As ethanol has detrimental effects on performance in tasks sensitive to hippocampal impairment [42], such as episodic memory formation [43] and spatial learning [44,45], and alters the function of place cells in the hippocampus [46,47], a structure which seems to encode cognitive maps with strong similarities to the SR [24,48–50], non-alcohol-naïve

participants might have learned that cognitive-map-based value estimates are often inaccurate in alcohol-related contexts and thus default to computationally cheaper MF behavior. Indeed, contextual cues paired with alcohol intoxication impaired goal-directed behavior in rodents [51] (but see [52]). Future research in individuals with AUD versus no prior drinking experience should further elucidate context-dependent strategy contributions.

In conclusion, we replicated a key signature of SR use [21] in a naturalistic task and showed that this signature can also indicate use of a rigid goal-based reduced version of SR [22]. However, low-risk drinkers' behavior in non-alcohol-related contexts did not align with canonical SR's policy dependence but instead with a hybrid (reduced) random-policy SR-MF strategy. To our knowledge, this is the first study to test associations between use of SR-based RL and psychopathology. We expected a shift away from a MB towards a (red)SR strategy in high-risk drinkers but instead found a shift away from a (reduced) random-policy SR-like strategy towards more habitual MF RL in alcohol-related contexts. Our results suggest that inconsistent evidence for increased habitual behavior in human addiction cannot be explained by enhanced use of SR-based strategies as more nuanced substrates of habit [21]. Instead, random-policy SR-based RL, even its goal-based dimension-reduced variant, is a resource-rational strategy enabling flexible behavior under a range of circumstances, rather than a computational proxy for pathologically inflexible behavior in addiction. That substance-related contexts shifted strategy use towards MF RL underscores the importance of considering context when testing hypotheses about altered cognitive map representations in mental health conditions.

Methods

Participants and procedures

Participants between 18 and 45 years old who did not generally abstain from or wish to abstain from alcohol at the time of study participation and had normal or corrected-to-normal vision were recruited via Prolific. All participants gave written informed consent and were compensated with £12 for 1.5h of time, plus an optional bonus payment (see multi-stage decision-making paradigm). The study was approved by the ethics committee at Dresden University of Technology, Germany (no. SR-EK-578122022).

Participants first answered basic demographic questions and completed the AUDIT questionnaire [28,29] which was used to determine group allocation. High-risk drinkers had an AUDIT score >7, whereas low-risk drinkers had an AUDIT score <8 and indicated that they had never been in individual therapy for alcohol use. Participants then completed the multi-stage

sequential decision-making task before filling in the AUDIT questionnaire a second time. We additionally assessed AUD criteria according to the Diagnostic and Statistical Manual of Mental Disorders (DSM)-5 [53] using a self-assessment questionnaire [54,55], drinking days, drinks per drinking day, and binge drinking days in the past 3 months via a Quantity Frequency Questionnaire according to the Munich Composite International Diagnostic Interview [56,57], automated drinking behavior via the Craving Automated Scale for Alcohol (CAS-A) [58], tobacco, cannabis, and other drug use in the past three months via custom screening questions, impulsivity via the Urgency-Premeditation-Perseverance-Sensation Seeking-Positive Urgency (UPPS-P) scale [59,60], and obsessive compulsive symptoms via the Obsessive Compulsive Inventory-Revised (OCI-R) [61,62].

To achieve the preregistered sample size of *N*=420 after an estimated exclusion rate of 25%, we initially recruited *N*=560 individuals with complete task and questionnaire data. After application of exclusion criteria (see below), we continued to recruit until we reached a full balancing of task condition order and an equal sample size in each of the four subsamples, resulting in recruitment of a total of *N*=668 participants with complete data (*n*=167 low-risk drinkers in non-alcohol context; *n*=158 low-risk drinkers in alcohol context; *n*=175 high-risk drinkers in non-alcohol context; *n*=168 high-risk drinkers in alcohol context). Out of these, *N*=54 (*n*=14 low-risk drinkers in non-alcohol context; *n*=6 low-risk drinkers in alcohol context; *n*=21 high-risk drinkers in non-alcohol context; *n*=13 high-risk drinkers in alcohol context) were excluded as their answers in the pre- and post-task AUDIT questionnaires indicated a different group allocation. To ensure that all participants had learned which path was optimal by the end of the learning phase, we further excluded *N*=94 individuals who had chosen the optimal path in less than 3 out of the last 5 learning trials in any task condition (*n*=23 low-risk drinkers in non-alcohol context; *n*=22 low-risk drinkers in alcohol context; *n*=24 high-risk drinkers in non-alcohol context; *n*=25 high-risk drinkers in alcohol context). Our final sample thus consisted of *N*=520 individuals (*n*=130 per subsample; see Table 1 for sample characteristics).

Multi-stage decision-making paradigm

The multi-stage decision-making task adapted from Momennejad et al. [21] consisted of five task conditions (reward revaluation, goal-state revaluation, transition revaluation, policy revaluation, control condition) encompassing four phases (learning, re-learning, test, rating) each (Figure 1A). The order of conditions was balanced in a Latin square design across participants. Participants performed the task in either an alcohol or a non-alcohol context, with each condition taking place in a separate environment. Which condition took place in which

environment was approximately balanced across participants (Figure S13). Images representing states in the different environments were generated using OpenAI's DALL-E (Figures 1B, S1; for stimulus material, see publicly available code). Which state was represented by which image in each environment was randomized within each stage of the decision-tree.

In the non-alcohol context, participants navigated through five apartment environments (Figure 1) consisting of 10 rooms (states), to collect as much hidden cash as possible. The cover story was that every Friday, they would go out for dinner with friends to a restaurant which accepted no card payments. They regularly forgot to withdraw cash during the week but luckily made it a habit to deposit some cash for emergencies in certain spots in their apartment. How much cash they deposited where, however, they could not remember. Shortly before going to the restaurant on Fridays, they would thus have to search the usual spots for cash. Changes in apartment environments between conditions were explained by staying in a vacation home or moving from time to time.

In the alcohol context, participants navigated through five bar environments (Figure 1) consisting of 10 rooms (states) each to collect as many glasses of their preferred alcoholic drink as possible. Participants could choose their preferred drink (beer, wine, long drinks/cocktails) before starting the task. Here, the cover story was that every Saturday after a stressful week at work, they would go for drinks in their favorite bar. As the bar was popular, some rooms would usually be booked for private events, and bar counters in some rooms would serve their favorite drink only in small quantities or be out of it entirely. While they would never know in what room they would get what number of glasses, they would try to collect as many as possible. Changes in bar environments between conditions were explained by having a new favorite bar from time to time.

In both contexts, participants were instructed that how much reward they received in a room would not depend on how they got there, that what had been learned in one environment would not be valid in others, that they could take a self-paced break between environments, and that they had to choose an action as quickly as possible in each room. If they passed the time limit (2s in learning and re-learning trial states with one available action, 3s in learning trial states with two available actions, 15s in test trial states; Figure S1), they would start the current trial from the beginning. Left actions were selected using the ‘F’ key, right actions using the ‘J’ key, and single actions using the ‘space’ key. As an incentive, participants were informed that the person who collected the most cash or glasses across the entire task would get a bonus payment (non-alcohol context) or a beverage store voucher (alcohol context) of £50, and that

participants in the second to fourth place would receive bonus payments or vouchers of £15 each. At the end of the study, participants in the alcohol context task were informed that, for ethical reasons, the most successful individuals would not receive beverage store vouchers, but instead bonus payments. We determined the four winners for each of the four subsamples (16 winners in total) and paid bonus payments via Prolific after completion of data collection.

Before starting the task, all participants read instructions, saw a floor plan of the environment, performed a 10-trial training version of the task, and completed a quiz. The floor plan did not show the underlying task structure depicted in Figure 1, but instead showed states on the same stage, e.g. states 4-6, as adjacent numbered squares without doors to not imply a specific transition structure. The training consisted of 4 forced-choice and 6 free-choice trials starting from state 1 and was performed in an environment ('light blue apartment' or 'tapas bar') not used on the actual task to avoid carry-over effects. The quiz encompassed 8 true/false questions about the instructions with no time limit. If any of the questions was answered incorrectly, participants restarted from the beginning.

Only once all quiz questions had been answered correctly, participants started the actual task. In the initial 24-trial learning phase per condition, all trials started from state 1 of the 10-state environment. States 1-3 contained two actions (doors) each and states 4-10 contained a single action (searching a spot in an apartment room or ordering at a bar counter), respectively. A path through the environment was thus determined by first- and second-stage actions. Rewards were delivered in fourth-stage states and implied that one path starting from state 1 was optimal, i.e. led to the highest reward. Which state-1-action initiated the optimal path was randomized between conditions (i.e. rewards in states 7 and 9 were swapped), with three conditions having an optimal path starting with the left state-1-action and two conditions having an optimal path starting with the right state-1-action. The first four learning trials were forced-choice, i.e. only one door in states 1-3 was available and highlighted and the other one was greyed out. In the following 20 free-choice trials, both doors in states 1-3 were available. Before starting the task, participants had been instructed to memorize during forced-choice trials where they would receive the most reward and then to collect as much reward as possible.

In the following 9-trial re-learning phase, participants experienced changes to the reward structure (in reward, goal-state, policy revaluation and the control condition) or transition structure (in transition revaluation) of the environment (for a detailed description, see Figure 1). Changes in all revaluation conditions implied a change in optimal path compared to what had been experienced during the learning phase, whereas changes in the control condition did not.

All re-learning trials started from third-stage states (3 from state 4, 5, and 6, each) and were presented in randomized order. Before beginning the task, participants had been instructed that they would sometimes start in later rooms than the first one, that sometimes things would change in an environment, and that they would have to notice these changes and act accordingly to collect as much reward as possible. Immediately before each condition's re-learning phase, participants were only informed that a few weeks had passed since they went out for dinner (non-alcohol context) or had been to their favorite pub (alcohol context) and that now, they would save themselves some time and start from later rooms.

In the 5-trial test phase per condition, participants started from state 1 again. Immediately before this phase, they were instructed that they would not see where their actions would take them and how much reward they would receive, but that they should nevertheless choose the path that they believed would lead them to the highest amount of reward and that their decisions would be considered for the bonus payments or vouchers. The test phase served to derive the primary measure of revaluation performance, i.e. to assess whether, on a given test trial, a participant chose a different path from the path which had been optimal during the learning phase.

In the final rating phase per condition, participants were presented two times with state 1, with one action being highlighted, respectively. On each of the two screens, they were asked how much cash or how many glasses of alcohol the highlighted action would lead to and could indicate their answer on a sliding scale ranging from 0 = 'no cash / glasses of alcohol' to 100 = 'lots of cash / glasses of alcohol' using the left and right arrow keys. Numbers were not shown on screen, and answers were confirmed using the 'Enter' key. This phase allowed to derive a secondary measure of revaluation performance, i.e. the difference in rating scores between state-1 actions.

Computational models and simulations

We simulated four *a priori* defined RL algorithms – a SR, a redSR, a MB, and a MF learner – in the multi-stage decision-making task to compare their behavior to that of low-risk and high-risk drinkers. As none of these models nor a combination of them could account for the observed behavior pattern in humans, we additionally simulated several *post hoc* defined models: a MB algorithm with different learning rates for the reward and transition structure, a random-policy SR model, and a reduced random-policy SR algorithm.

Environment and policy

The environment's transition and reward structure (monetary rewards) were hard-coded, with 10 states and 13 state-action pairs in total (two actions from states 1-3, respectively). Each simulation encompassed 4 forced-choice learning trials in random order, 20 free-choice learning trials, and 9 re-learning trials.

Q-values derived from all models were transformed into action probabilities using a standard softmax policy π without temperature parameter:

$$\pi(a|s) = \frac{e^{Q_{sa}}}{\sum_{a'} e^{Q_{sa'}}} \quad [1]$$

, where $Q_{sa'}$ denotes the Q-values of all actions available in state s.

SR algorithm

The SR algorithm [17,18,20] represents the environment via a state-action-state-action successor matrix M^{π}, where each state-action-pair sa (in row sa) is represented by the expected cumulative future discounted number of visits to each successor state-action-pair $s'a'$ (in column $s'a'$), when starting from sa and following the softmax policy π:

$$M^{\pi}_{sa,s'a'} = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \, \mathbb{I}(sa_t = s'a') | sa_0 = sa\right] \quad [2]$$

Here, $\mathbb{I}(\cdot)$ equals 1 if its argument is true, i.e. if the agent transitioned to $s'a'$ starting from sa in the current time step t, and 0 otherwise, and γ is a time discount factor between 0 and 1, with higher values indicating less discounting.

The successor matrix M^{π} is, in our case, learned from experience using temporal difference (TD) updates. Specifically, after transitioning from a state-action-pair sa to a successor state-action-pair $s'a'$, the row of M^{π} corresponding to sa is updated as follows:

$$M^{\pi}_{sa,:} \leftarrow M^{\pi}_{sa,:} + \alpha_{SR}\left[\mathrm{I}_{sa} + \gamma M^{\pi}_{s'a',:} + M^{\pi}_{sa,:}\right] \quad [3]$$

Here, α_{SR} is a SR learning rate between 0 and 1, and I_{sa} is a vector containing zeros at each index expect for a 1 at the index corresponding to the state-action-pair sa which the agent transitioned from.

To derive a value estimate for action a from state s, the SR algorithm uses the row of M^{π} corresponding to sa as a feature vector and linearly weights it by the elements in a vector w:

$$Q_{sa} = \sum_{s'a'} M^{\pi}_{sa,s'a'} w_{s'a'} \quad [4]$$

The weight vector w is itself learned from experience via TD-learning optimized for linear function approximation instead of delta-rule learning of immediate punctate rewards [18,20,22] and gradually approaches the one-step reward received in each state-action pair sa. Specifically, after a transition $sa \rightarrow s'a'$, all elements i of w are updated as follows:

$$w_i \leftarrow w_i + \alpha_{TD} \delta_{RPE} M^{\pi}_{sa,i} \quad [5]$$

, where α_{TD} is a TD learning rate, the feature vector $M^{\pi}_{sa,:}$ is scaled by $M^{\pi}_{sa,:} \times {M^{\pi}_{sa,:}}^{T}$ to allow for interpretation of α_{TD} as a proportional step-size [20], and δ is the TD reward prediction error:

$$\delta_{RPE} = R_{sa} + \gamma Q_{s'a'} - Q_{sa} \quad [6]$$

In our implementation of the SR algorithm, M^{π} was initialized as a 13×13 identity matrix and w and Q as all-zero vectors, respectively. Upon entering a state s, the agent decided for an action a using equation [1] and, upon receiving the corresponding reward, first updated the weight vector w using equation [5] and then updated Q-values using equation [4] based on the updated w and the next predicted action $s'a'$. After transitioning to the next state s', the agent then updated the successor matrix M^{π} using equation [3].

RedSR algorithm

In learning phase trials, the redSR algorithm [22] operates exactly like the SR learner. Before entering the re-learning phase, M^{π} is then reduced to M^{π}_{red} which contains columns only for goal-state-action-pairs, i.e. state-action-pairs for which the corresponding element in w is > 1 (to disregard slight numerical deviations around zero). The weight vector w is also reduced to w_{red} containing elements only for goal-state-action-pairs. During re-learning, M^{π}_{red} is rigid, i.e. not updated, whereas w_{red} continues to be updated according to equation [5].

The fact that all elements of w_{red} are updated after each transition, and that the update is based on the dimension-reduced, inaccurate M^{π}_{red}, is what has previously been reported to cause redSR's addiction-like characteristics, such as sustained positive RPEs upon reaching goal-states under a resistant policy [22]. These characteristics would not be observed if redSR would use delta-rule learning of immediate punctate rewards [20,22]. However, it is also what causes redSR to succeed in goal-state revaluation under certain parameter settings (Figure S1). Specifically, in goal-state revaluation, the reward prediction error δ_{RPE} caused by the new reward updates all elements i of w_{red} which correspond to old goal-states that have previously been reached from the new goal-state (old goal-states in column i are > 0 in $M^{\pi}_{sa,i}$, see equation

[5]). While w_{red} does not contain an element representing the new goal-state directly, the elements i of w_{red} corresponding to old goal-states are thus inflated and affect value estimation according to equation [4]. An alternative version of goal-state revaluation where the new goal is introduced after a previous one would avoid this. Here, we decided against this alternative version to exclude that failure to adapt to goal-state revaluation might simply be due to scale dependency, i.e. to individuals representing the environment only up until previous goal-states.

As there is no definitive answer to the question which state-action-pairs are actually perceived as goal-state-action-pairs, we implemented two alternative versions of redSR which additionally consider the final $s10a1$ ('3 goals' in Figure S6) or $s10a1$ and the third state-action-pair $s7a1$, $s8a1$, or $s9a1$ on the same stage as the previously reward-containing state-action-pairs ('4 goals' in Figure S6) as additional goal-state-action-pairs.

MB algorithm

The MB learner [63] represents the environment by a one-step state-action-state transition matrix T, where each element $T_{sa,s'}$ equals the probability of reaching state s' (in columns s') in the immediate next timestep t when performing action a in state s (in row sa):

$$T_{sa,s'} = P(s_{t+1} = s' | s_t = s, a_t = a) \quad [7]$$

T is learned from experience by counting observed transitions $sa \rightarrow s'$ and updated according to the following rule:

$$T_{sa,s'} \leftarrow T_{sa,s'} + 1 \quad [8]$$

Elements $T_{sa,s'}$ are then normalized to ensure that each row $T_{sa,:}$ adds up to 1:

$$T_{sa,s'} = \frac{T_{sa,s'}}{\sum_{s''} T_{sa,s''}} \quad [9]$$

Here, s'' denotes all states in the environment.

To derive Q-value estimates for a state-action-pair sa, the MB learner iterates the Bellman equation, which combines T with knowledge about the environment's reward structure R, until convergence:

$$Q^{\pi}_{sa} = R_{sa} + \gamma \sum_{s'} T_{sa,s'} \sum_{a'} \pi(a'|s') Q^{\pi}_{s'a'} \quad [10]$$

One-step rewards R_{sa} are themselves learned from experience using TD updates following each transition $sa \rightarrow s'$:

$$R_{sa} \leftarrow R_{sa} + \alpha_{TD}\delta_{RPE} \quad [11]$$

, where δ_{RPE} is defined as in equation [6].

T was initialized as a 13×10 matrix of all zeros, and R and Q as all-zero vectors. In our implementation, the agent decided for an action a upon entering a state s and then updated the reward vector R using equation [11] based on the reward just observed and the next predicted action $s'a'$ before re-estimated all Q-values using equation [10]. After transitioning to s', T was updated using equation [8].

MF algorithm

The MF learner [63] has no representation of the environment. Instead, it caches Q-values directly and updates them after each transition $sa \rightarrow s'a'$ using TD learning:

$$Q_{sa} \leftarrow Q_{sa} + \alpha_{TD}\delta_{RPE} \quad [12]$$

, where δ_{RPE} is defined as in equation [6]. We initialized Q as an all-zero vector.

MB algorithm with separate learning rates for T and R

The MB algorithm with a TD learning rate α_{TD} for the reward structure and a learning rate α_T for the transition structure [64,65] operates like the MB learner, with the exception that T is learned via state prediction error δ_{SPE}-based updates. Specifically, following a transition $sa \rightarrow s'a'$, the row of T corresponding to sa is updated as follows:

$$T_{sa,s''} \leftarrow T_{sa,s''} + \alpha_T\delta_{SPE} \quad [13]$$

, where s'' denotes all states of the environment and δ_{SPE} is defined as

$$\delta_{SPE} = \mathrm{I}_{s'} - T_{sa,s''} \quad [14]$$

$\mathrm{I}_{s'}$ is a vector of zeros except for a 1 at the index corresponding to the state s' which the agent transitioned to.

We initialized T as a 13×10 matrix with a small non-zero prior for all transitions, i.e. all elements had a value of 1 / *n* states = 0.1. R and Q were initialized as all-zero vectors. Upon entering a state s, the agent decided for an action a and, after observing the corresponding reward, updated the reward vector R using equation [11] based on the reward and the next predicted

action $s'a'$ before updating Q-values using equation [10]. After transitioning to s', T was updated using equation [13].

Random-policy SR algorithm

The random-policy SR algorithm learned M^{random} during the initial forced-choice trials. While this can happen in several ways (see Discussion), here, we implemented the standard MB learner described above to learn T using equations [8] and [9] and R using equation [11]. After trial 4, i.e. before entering the first free-choice learning trial, the state-action-state transition matrix T is first transformed to a state-action-state-*action* transition matrix under a random policy T^{random}:

$$T_{sa,s'a'} = T_{sa,s'} / \sum s'a'' \quad [15]$$

, with $\sum s'a''$ being the number of available actions in state s'.

Each element $T_{sa,s'a'}$ equals the probability of performing action a' in state s' (in columns $s'a'$) in the immediate next timestep t when performing action a in state s (in row sa):

$$T_{sa,s'a'} = P(s_{t+1} = s', a_{t+1} = a' | s_t = s, a_t = a) \quad [16]$$

T^{random} is then transformed to M^{random} using matrix inversion [20]:

$$M^{random} = (I - \gamma T^{random})^{-1} \quad [17]$$

The weight vector w used is SR is set to equal R.

Starting with the first free-choice learning trial, the random-policy SR algorithm then uses equation [4] to compute value estimates. However, M^{random} and w are rigid, i.e. not updated, during the learning phase. Only during re-learning trials, the random-policy SR algorithm operates exactly like the SR algorithm described above, using equation [3] to update M^{random} and equation [5] to update w.

As M^{random} is the defining feature of random-policy SR, we implemented an alternative version of the random-policy SR algorithm where w is updated throughout free-choice learning phase trials ('rigid M' in Figure S5).

Reduced random-policy SR algorithm

The reduced random-policy SR algorithm operated like the alternative random-policy SR algorithm where w (equation [5]), but not M^{random}, is updated throughout free-choice learning

phase trials, with the exception that, before entering the re-learning phase, M^{random} was reduced to M_{red}^{random} containing columns only for goal-state-action-pairs, and w was reduced to w_{red} containing elements only for goal-state-action-pairs. During re-learning, M_{red}^{random} remained rigid, whereas w_{red} was continually updated using equation [5].

Like for redSR, we simulated three versions of reduced random-policy SR with different numbers of goal-state-action-pairs ('2 goals', '3 goals', and '4 goals' in Figure S6). Figure 3C depicts the version with 4 goals, i.e. the state-action-pairs for which the corresponding element in w is > 1 after learning (to disregard slight numerical deviations around zero), the third state-action pair $s7a1$, $s8a1$, or $s9a1$ on the same stage as the previously reward-containing state-action-pairs, and the final $s10a1$.

Simulation parameters and performance evaluation

All models, except the MB learner with different learning rates for R and T, had two free parameters: a learning rate α and a time discount factor γ. For SR, redSR, random-policy SR, and reduced random-policy SR algorithms, this means that α_{TD} and α_{SR} were assumed to be equal. We performed simulations at all possible combinations of 0.5, 0.7, and 0.9 for α and γ, respectively. A learning rate lower than 0.5 was considered implausible as it would not allow successful learning and re-learning in the complex task environment within relatively few trials. A γ lower than 0.5 would indicate strong discounting which seemed implausible given the relatively short time horizon of our task. The MB learner with different learning rates for R and T was simulated at α_{TD}=0.9 and α_{T}=[0.1, 0.3, 0.5]. Parameters were assumed to stay constant over the course of all task phases.

We conducted 1000 simulation runs per model and condition. After the learning and re-learning phase, respectively, we conducted one test trial to check whether agents had acquired a higher value for the optimal first state and second-stage choice. Agents were considered to have acquired an optimal path preference if they were successful in the learning test trial and if the optimal path additionally had been chosen in at least 3 out of the last 5 learning trials (same criterion as applied to human data). Only successful learning runs were considered to determine revaluation performance. Agents were considered to have successfully changed their path preference if they had acquired a higher value for the now optimal first state and second-stage choice in the re-learning test trial. For example, if an agent had learned successfully in 500 runs in a certain condition, a successful change in path preference after re-learning in 250 of these runs would indicate a revaluation performance of 50%.

Mixed-effects regression analyses of revaluation performance

Mixed-effects regression models were implemented using the *lme4* package [66] in *R* version 4.4.1 [67]. We included the maximum random effect structure supported by Likelihood Ratio Tests. Significance tests and confidence intervals for fixed effects were based on Wald degrees of freedom for logistic mixed-effects models, and on Satterthwaite's degrees of freedom for linear-mixed-effects models. We performed post-hoc tests for significant interactions using the *emmeans* package [68] and report Bonferroni-corrected *p*-values. Analyses were based on complete trials only, i.e. trials were excluded if a participant exceeded a state's response time limit.

Hypothesis tests

To test the preregistered hypotheses, we used logistic mixed-effects regression models predicting revaluation performance on a given test trial, i.e. the probability of choosing a different path from the one which had been optimal during the learning phase. For the four revaluation conditions, this variable was coded as 1 if the path chosen on a test trial was optimal according to what had been experienced during re-learning, and as 0 otherwise. Conversely, for the control condition, this variable was coded as 1 if any but the optimal path according to what had been experienced during learning and re-learning was chosen erroneously, and as 0 if the optimal path was chosen.

To test hypotheses 1 and 2 in low-risk drinkers performing the task in the non-alcohol context, we included a fixed effect for dummy-coded condition with transition revaluation (hypothesis 1) or policy revaluation (hypothesis 2) as reference category, respectively. The model included a by-subject random intercept and random slope for condition to control for individual differences in revaluation performance in the reference condition as well as in contrasts between conditions of interest and the reference condition. We used the following *lme4* code:

```
change_path ~ condition + (1 + condition | participantID)
```

To test hypotheses 3, 4, and 6 in the full sample, we added fixed effects for group and context as well as all possible 2-way and 3-way interactions and kept the same random effect structure as above. Condition was dummy-coded with transition revaluation (hypothesis 3 and 4) and reward revaluation (hypothesis 6) as reference category, respectively. Group (-0.5=low-risk, 0.5=high-risk) and context (-0.5=non-alcohol, 0.5=alcohol) were sum-contrast-coded in all analyses. We used the following *lme4* code:

```
change_path ~ group*context*condition + (1 + condition |
participantID)
```

To test hypothesis 5 in low-risk and high-risk drinkers performing the task in an alcohol context, revaluation performance on a given test trial was predicted from fixed effects for dummy-coded condition with reward revaluation as a reference category, group, and their interaction. We again used the same random-effects structure as above. The *lme4* code used thus was as follows:

```
change_path ~ group*condition + (1 + condition | participantID)
```

In low-risk drinkers performing the non-alcohol task version and in the full sample, we repeated models with the control condition as a reference category to exclude random behavior or forgetting as reasons for revaluation success and with reward or policy revaluation as reference categories, respectively, to support our argument.

Individual slopes for the different contrasts of the condition predictor were derived by adding the participant's random slope extracted using *lme4*'s *ranef()* function to individually applicable fixed effects. Correlations amongst slopes or with other measures were quantified via Spearman's rank correlation coefficient.

Exploratory analyses of rating phase data

We explored performance in the rating phase using linear mixed-effects regression. To derive the rating score difference per participant as a dependent variable, we subtracted the rating of the state-1 action which was optimal during learning from the rating of the state-1 action which was suboptimal during learning. Positive rating score differences thus indicate a preference for the state-1 action which was suboptimal during learning and optimal (in revaluation conditions) or still suboptimal (in control condition) according to what had been experienced during re-learning.

Analogously to hypotheses 1 and 2, we explored condition effects in low-risk drinkers performing the non-alcohol task version using the following *lme4* code:

```
diff_rating_score ~ condition + (1 | participantID)
```

Analogously to hypotheses 3 and 5, we explored group, context, and condition effects in the full sample using *lme4* code as follows:

```
diff_rating_score ~ group*context*condition + (1 | participantID)
```

Analogously to hypothesis 5, we explored group and condition effects in the alcohol context using the following *lme4* code:

```
diff_rating_score ~ group*condition + (1 | participantID)
```

Models did not include a by-subject random slope for condition, as each participant had only one rating score difference value. Predictors were contrast-coded as described for confirmatory logistic mixed-effects regression models.

Preregistration, data, and code availability

This study was preregistered prior to data collection at doi.org/10.17605/OSF.IO/9TUZE. For explanations regarding deviations from the preregistration, see Supplementary Material. The data are available from the corresponding author upon request. Code used to produce the results described in this manuscript is available at https://github.com/agschlagenhauf/SR_in_AUD_behav.git.

References

1. Watson, P., O'Callaghan, C., Perkes, I., Bradfield, L. & Turner, K. Making habits measurable beyond what they are not: A focus on associative dual-process models. *Neuroscience & Biobehavioral Reviews* **142**, 104869 (2022).
2. Vandaele, Y. & Ahmed, S. H. Habit, choice, and addiction. *Neuropsychopharmacology* **46**, 689–698 (2021).
3. Du, Y., Krakauer, J. W. & Haith, A. M. The relationship between habits and motor skills in humans. *Trends in Cognitive Sciences* **26**, 371–387 (2022).
4. Corbit, L. H. & Janak, P. H. Habitual alcohol seeking: Neural bases and possible relations to alcohol use disorders. *Alcoholism: Clinical and Experimental Research* **40**, 1380–1389 (2016).
5. Ersche, K. D. *et al.* Carrots and sticks fail to change behavior in cocaine addiction. *Science* **352**, 1468–1471 (2016).
6. Everitt, B. J. & Robbins, T. W. Drug addiction: Updating actions to habits to compulsions ten years on. *Annu. Rev. Psychol.* **67**, 23–50 (2016).
7. Lüscher, C., Robbins, T. W. & Everitt, B. J. The transition to compulsion in addiction. *Nat Rev Neurosci* **21**, 247–263 (2020).
8. Voon, V. *et al.* Disorders of compulsivity: A common bias towards learning habits. *Mol Psychiatry* **20**, 345–352 (2015).
9. Doñamayor, N. *et al.* Goal-directed and habitual control in human substance use: state of the art and future directions. *NPS* **81**, 403–417 (2022).
10. Vandaele, Y. & Janak, P. H. Defining the place of habit in substance use disorders. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* **87**, 22–32 (2018).
11. Hogarth, L. Addiction is driven by excessive goal-directed drug choice under negative affect: Translational critique of habit and compulsion theory. *Neuropsychopharmacology* **45**, 720–735 (2020).
12. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204–1215 (2011).
13. Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A. & Daw, N. D. Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *eLife* **5**, e11305 (2016).
14. Otto, A. R., Raio, C. M., Chiang, A., Phelps, E. A. & Daw, N. D. Working-memory capacity protects model-based learning from stress. *Proceedings of the National Academy of Sciences* **110**, 20941–20946 (2013).
15. Miller, K. J., Shenhav, A. & Ludvig, E. A. Habits without values. *Psychol Rev* **126**, 292–311 (2019).
16. Collins, A. G. E. & Cockburn, J. Beyond dichotomies in reinforcement learning. *Nature Reviews Neuroscience* **21**, 576–586 (2020).
17. Dayan, P. Improving Generalization for Temporal Difference Learning: The Successor Representation. *Neural Computation* **5**, 613–624 (1993).

18. Gershman, S. J. The Successor Representation: Its Computational Logic and Neural Substrates. *J. Neurosci.* **38**, 7193–7200 (2018).

19. Carvalho, W., Tomov, M. S., de Cothi, W., Barry, C. & Gershman, S. J. Predictive representations: building blocks of intelligence. Preprint at http://arxiv.org/abs/2402.06590 (2024).

20. Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J. & Daw, N. D. Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLOS Computational Biology* **13**, e1005768 (2017).

21. Momennejad, I. *et al.* The successor representation in human reinforcement learning. *Nat Hum Behav* **1**, 680–692 (2017).

22. Shimomura, K., Kato, A. & Morita, K. Rigid reduced successor representation as a potential mechanism for addiction. *European Journal of Neuroscience* **53**, 3768–3790 (2021).

23. Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J. & Daw, N. D. Neural evidence for the successor representation in choice evaluation. 2021.08.29.458114 Preprint at https://doi.org/10.1101/2021.08.29.458114 (2021).

24. Garvert, M. M., Saanum, T., Schulz, E., Schuck, N. W. & Doeller, C. F. Hippocampal spatio-predictive cognitive maps adaptively guide reward generalization. *Nat Neurosci* **26**, 615–626 (2023).

25. Ekman, M., Kusch, S. & Lange, F. P. de. Successor-like representation guides the prediction of future events in human visual cortex and hippocampus. *eLife* https://elifesciences.org/articles/78904 (2023) doi:10.7554/eLife.78904.

26. Kahn, A. E. & Daw, N. D. Humans rationally balance detailed and temporally abstract world models. *Communications Psychology* **3**, 1 (2025).

27. Redish, A. D. Addiction as a Computational Process Gone Awry. *Science* **306**, 1944–1947 (2004).

28. Saunders, J. B., Aasland, O. G., Babor, T. F., De La Fuente, J. R. & Grant, M. Development of the Alcohol Use Disorders Identification Test (AUDIT): WHO Collaborative Project on Early Detection of Persons with Harmful Alcohol Consumption-II. *Addiction* **88**, 791–804 (1993).

29. Fragebogeninstrumente. *alkoholleilinie.de* https://alkoholleitlinie.de/diagnostik/fragebogen-instrumente/.

30. Piray, P. & Daw, N. D. Reconciling Flexibility and Efficiency: Medial Entorhinal Cortex Represents a Compositional Cognitive Map. 2024.05.16.594459 Preprint at https://doi.org/10.1101/2024.05.16.594459 (2024).

31. Piray, P. & Daw, N. D. Linear reinforcement learning in planning, grid fields, and cognitive control. *Nat Commun* **12**, 4942 (2021).

32. Wientjes, S. & Holroyd, C. B. The successor representation subserves hierarchical abstraction for goal-directed behavior. *PLOS Computational Biology* **20**, e1011312 (2024).

33. Geerts, J. P., Gershman, S. J., Burgess, N. & Stachenfeld, K. L. A probabilistic successor representation for context-dependent learning. *Psychological Review* (2023) doi:10.1037/rev0000414.

34. Otto, A. R., Gershman, S. J., Markman, A. B. & Daw, N. D. The Curse of Planning: Dissecting Multiple Reinforcement-Learning Systems by Taxing the Central Executive. *Psychol Sci* **24**, 751–761 (2013).

35. de Cothi, W. *et al.* Predictive maps in rats and humans for spatial navigation. *Current Biology* **32**, 3676-3689.e5 (2022).

36. Yang, Y., Stöckl, C. & Maass, W. A surprising link between cognitive maps, successor-relation based reinforcement learning, and BTSP. 2025.04.22.650046 Preprint at https://doi.org/10.1101/2025.04.22.650046 (2025).

37. Bittner, K. C., Milstein, A. D., Grienberger, C., Romani, S. & Magee, J. C. Behavioral time scale synaptic plasticity underlies CA1 place fields. *Science* **357**, 1033–1036 (2017).

38. Wu, Y. & Maass, W. A simple model for Behavioral Time Scale Synaptic Plasticity (BTSP) provides content addressable memory with binary synapses and one-shot learning. *Nat Commun* **16**, 342 (2025).

39. John, T. *et al.* Representation of visual sequences in the tuning and topology of neuronal activity in the human hippocampus. 2025.03.04.641300 Preprint at https://doi.org/10.1101/2025.03.04.641300 (2025).

40. George, T. M., de Cothi, W., Stachenfeld, K. L. & Barry, C. Rapid learning of predictive maps with STDP and theta phase precession. *eLife* **12**, e80663 (2023).

41. Schwöbel, S., Marković, D., Smolka, M. N. & Kiebel, S. J. Balancing control: A Bayesian interpretation of habitual and goal-directed behavior. *Journal of Mathematical Psychology* **100**, 102472 (2021).

42. White, A. M., Matthews, D. B. & Best, P. J. Ethanol, memory, and hippocampal function: A review of recent findings. *Hippocampus* **10**, 88–93 (2000).

43. Söderlund, H., Grady, C. L., Easdon, C. & Tulving, E. Acute effects of alcohol on neural correlates of episodic memory encoding. *NeuroImage* **35**, 928–939 (2007).

44. Matthews, D. B., Simson, P. E. & Best, P. J. Acute ethanol impairs spatial memory but not stimulus/response memory in the rat. *Alcohol Clin Exp Res* **19**, 902–909 (1995).

45. Matthews, D. B., Ilgen, M., White, A. M. & Best, P. J. Acute ethanol administration impairs spatial performance while facilitating nonspatial performance in rats. *Neurobiol Learn Mem* **72**, 169–179 (1999).

46. Matthews, D. B., Simson, P. E. & Best, P. J. Ethanol Alters Spatial Processing of Hippocampal Place Cells: A Mechanism for Impaired Navigation When Intoxicated. *Alcoholism: Clinical and Experimental Research* **20**, 404–407 (1996).

47. Miyake, K. *et al.* Acute Effects of Ethanol on Hippocampal Spatial Representation and Offline Reactivation. *Frontiers in Cellular Neuroscience* **Volume 14-2020**, (2020).

48. Stachenfeld, K. L., Botvinick, M. M. & Gershman, S. J. The hippocampus as a predictive map. *Nat Neurosci* **20**, 1643–1653 (2017).

49. Cothi, W. de & Barry, C. Neurobiological successor features for spatial navigation. *Hippocampus* **30**, 1347–1355 (2020).

50. Brunec, I. K. & Momennejad, I. Predictive Representations in Hippocampal and Prefrontal Hierarchies. *J. Neurosci.* **42**, 299–312 (2022).

51. Ostlund, S. B., Maidment, N. T. & Balleine, B. W. Alcohol-Paired Contextual Cues Produce an Immediate and Selective Loss of Goal-directed Action in Rats. *Frontiers in Integrative Neuroscience* **Volume 4-2010**, (2010).

52. Sjoerds, Z. *et al.* Behavioral and neuroimaging evidence for overreliance on habit learning in alcohol-dependent patients. *Transl Psychiatry* **3**, e337–e337 (2013).

53. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*. (American Psychiatric Association, Washington, DC, 2013).

54. Alcohol Use Disorder: A Comparison Between DSM–IV and DSM–5. *National Institute on Alcohol Abuse and Alcoholism (NIAAA)* https://www.niaaa.nih.gov/publications/brochures-and-fact-sheets/alcohol-use-disorder-comparison-between-dsm.

55. Hildebrandt, M. K., Dieterich, R. & Endrass, T. Disentangling substance use and related problems: urgency predicts substance-related problems beyond the degree of use. *BMC Psychiatry* **21**, 242 (2021).

56. Kuitunen-Paul, S. *et al.* Assessment of alcoholic standard drinks using the Munich composite international diagnostic interview (M-CIDI): An evaluation and subsequent revision. *International Journal of Methods in Psychiatric Research* **26**, e1563 (2017).

57. Lachner, G. *et al.* Structure, Content and Reliability of the Munich-Composite International Diagnostic Interview (M-CIDI) Substance Use Sections. *European Addiction Research* **4**, 28–41 (1998).

58. Vollstädt-Klein, S., Leménager, T., Jorde, A., Kiefer, F. & Nakovics, H. Development and Validation of the Craving Automated Scale for Alcohol. *Alcoholism: Clinical and Experimental Research* **39**, 333–342 (2015).

59. Cyders, M. A. *et al.* Integration of impulsivity and positive mood to predict risky behavior: Development and validation of a measure of positive urgency. *Psychological Assessment* **19**, 107–118 (2007).

60. Wüllhorst, V., Lützkendorf, J. & Endrass, T. Validation of the German long and short versions of the UPPS-P Impulsive Behavior Scale. *Journal of Clinical Psychology* **80**, 2099–2116 (2024).

61. Foa, E. B. *et al.* The Obsessive-Compulsive Inventory: Development and validation of a short version. *Psychological Assessment* **14**, 485–496 (2002).

62. Gönner, S., Leonhart, R. & Ecker, W. Das Zwangsinventar OCI-R - die deutsche Version des Obsessive-Compulsive Inventory-Revised. *Psychother Psychosom Med Psychol* **57**, 395–404 (2007).

63. Sutton, R. S. & Barto, A. *Reinforcement Learning: An Introduction*. (The MIT Press, Cambridge, Massachusetts London, England, 2020).

64. Gläscher, J., Daw, N., Dayan, P. & O'Doherty, J. P. States versus Rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* **66**, 585–595 (2010).

65. Möhring, L. & Gläscher, J. Prediction errors drive dynamic changes in neural patterns that guide behavior. *Cell Reports* **42**, 112931 (2023).

66. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67**, 1–48 (2015).

67. R Core Team. R: A Language and environment for statistical computing. R Foundation for Statistical Computing (2024).

68. Lenth, R. V. Estimated Marginal Means, aka Least-Squares Means. (2024).

Acknowledgements

This project was supported by the Collaborative Research Center (CRC) 265 (project B01), funded by the German Research Foundation (grant no: 402170461 awarded to F.S. and T.E.). S.H.-M. was supported by a Philip Wrightson Fellowship from the New Zealand Neurological Foundation.

We would like to thank Christian Bator (github.com/christianbator) for his contributions to the HTML/JS/CSS code for the multi-stage decision-making task. We would also like to thank Claire Sturgill (torbenottlab.org/Claire Sturgill) for contributing to an initial version of simulations for *a priori* defined models and Ida Momennejad (momen-nejad.org) and Evan Russek (evanrussek.com) for helpful feedback on task design.

Author contributions

M.P.M.M., S.H-M., K.S., A.K., E.L.B., C.G., C.E., K.M., T.E., and F.S. designed the experiment. M.P.M.M. and K.S. ran model simulations. M.P.M.M. collected and analyzed the data. M.P.M.M. drafted the manuscript and S.H-M., K.S., A.K., E.L.B., C.G., C.E., K.M., S.J.G., T.E., and F.S. substantially revised it.

Competing interests

The authors declare no competing interests.

Table 1. Sample characteristics.

	Low-risk drinkers – monetary version (*n*=130)	**Low-risk drinkers - alcohol version (*n*=130)**	**High-risk drinkers – monetary version (*n*=130)**	**High-risk drinkers – alcohol version (*n*=130)**	**Group*version differences**		**Group differences**		**Version differences**	
					Test statistic	***p***	**Test statistic**	***p***	**Test statistic**	***p***
Age (*M* [*SD*])	29.82 (7.23)	31.01 (7.41)	30.94 (7.48)	31.72 (7.23)	t(516)=-0.32	.747	t(516)=1.23	.218	t(516)=1.31	.191
Sex (*n* male / female / other)	60/70/0	67/63/0	86/42/2	84/46/0	χ^2=22.33	<.001	χ^2=17.39	<.001	χ^2=2.12	.465
Country of residence (*n* GER / other EU country / UK)	53/10/67	63/0/67	14/21/95	39/0/91	χ^2(6)=79.42	<.001	χ^2(2)=35.34	<.001	χ^2(2)=38.30	<.001
Native language (n German / English / other)	44/64/22	52/59/19	19/88/20	29/87/11	χ^2(6)=31.38	<.001	χ^2(2)=26.40	<.001	χ^2(2)=4.37	.112
Ethnicity (*n* asian / black / white / mixed / other)	12/6/98/9/5	11/8/103/7/1	4/3/111/6/3	5/5/109/6/0	χ^2=16.85	.170	χ^2=10.07	.030	χ^2=6.33	.184
Student status (*n* student / no student)	47/67	39/76	37/79	34/80	χ^2(3)=3.75	.290	χ^2(1)=1.99	.158	χ^2(1)=0.90	.342
Employment status (*n* employed / job-seeking / not in paid work / other)	73/15/6/11	75/15/9/12	82/13/8/7	89/10/6/5	χ^2(9)=7.71	.564	χ^2(3)=6.04	.110	χ^2(3)=0.40	.939

	Low-risk drinkers – monetary version (*n*=130)	Low-risk drinkers - alcohol version (*n*=130)	High-risk drinkers – monetary version (*n*=130)	High-risk drinkers – alcohol version (*n*=130)	Group*version differences		Group differences		Version differences	
					Test statistic	*p*	Test statistic	*p*	Test statistic	*p*
AUDIT score (M [SD])	3.54 (1.65)	4.28 (1.79)	15.34 (5.27)	13.73 (5.08)	t(516)=-3.47	.001	t(516)=24.67	<.001	t(516)=1.54	.123
AUD criteria (M [SD])	0.62 (1.15)	0.92 (1.26)	4.85 (2.47)	4.15 (2.42)	t(516)=-2.98	.003	t(516)=17.69	<.001	t(516)=1.25	.210
AUD diagnosis (*n* fulfilled / not fulfilled)	16/114	32/98	117/13	111/19	$\chi^2(3)=254.96$	<.001	$\chi^2(1)=247.41$	<.001	$\chi^2(1)=0.63$	.429
Drinking days past 3 months (M [SD])	8.14 (7.03)[a]	12.30 (13.20)[d]	33.27 (22.64)[e]	27.23 (21.23)[g]	t(420)=-3.12	.002	t(420)=11.55	<.001	t(420)=1.98	.048
Drinks per drinking day in past 3 months (M [SD])	4.65 (5.79)[a]	4.95 (4.59)[d]	12.02 (10.52)[e]	12.11 (9.82)[g]	t(420)=-0.13	.896	t(420)=7.19	<.001	t(420)=0.30	.761
Binge days past 3 months (M [SD])	1.02 (2.76)[b]	2.21 (5.89)	16.86 (16.35)[f]	15.12 (16.18)[f]	t(509)=-1.39	.164	t(509)=10.64	<.001	t(509)=0.80	.423
Smoking past 3 months (*n* yes / no)	21/109	18/112	58/72	61/69	$\chi^2(3)=58.51$	<.001	$\chi^2(1)=56.74$	<.001	$\chi^2(1)=0.00$	1.00
Cannabis use past 3 months (*n* yes / no)	10/120	16/114	38/92	29/101	$\chi^2(3)=25.08$	<.001	$\chi^2(1)=20.95$	<.001	$\chi^2(1)=0.05$	.819
Other drug use past 3 months (*n* yes / no)	2/128	2/128	15/114	16/114	$\chi^2(3)=22.49$	<.001	$\chi^2(1)=20.82$	<.001	$\chi^2(1)=0.00$	1.00
CAS-A nonvolitional (M [SD])	0.41 (0.95)	0.41 (0.90)	3.11 (2.40)	2.86 (2.27)	t(516)=-0.79	.430	t(516)=12.24	<.001	t(516)=0.00	1.00
CAS-A unaware (M [SD])	0.33 (0.82)	0.34 (0.83)	3.19 (3.09)	2.79 (2.72)	t(516)=-1.09	.278	t(516)=10.79	<.001	t(516)=0.03	0.977

	Low-risk drinkers – monetary version (*n*=130)	Low-risk drinkers - alcohol version (*n*=130)	High-risk drinkers – monetary version (*n*=130)	High-risk drinkers – alcohol version (*n*=130)	Group*version differences		Group differences		Version differences	
					Test statistic	*p*	Test statistic	*p*	Test statistic	*p*
UPPS-P negative urgency (M [SD])	2.05 (0.57)	2.00 (0.63)	2.56 (0.58)	2.48 (0.56)	t(516)=-0.37	.709	t(516)=7.05	<.001	t(516)=-0.66	.509
UPPS-P premeditation (M [SD])	1.78 (0.47)	1.77 (0.44)	2.07 (0.48)	1.96 (0.44)	t(516)=-1.18	.239	t(516)=5.14	<.001	t(516)=-0.30	.767
UPPS-P perseverance (M [SD])	2.03 (0.59)	1.97 (0.55)	2.22 (0.54)	2.09 (0.54)	t(516)=-0.74	.459	t(516)=2.77	.006	t(516)=-0.80	.422
UPPS-P sensation seeking (M [SD])	2.34 (0.56)	2.37 (0.61)	2.58 (0.66)	2.51 (0.54)	t(516)=-0.93	.353	t(516)=3.30	.001	t(516)=0.41	.685
UPPS-P positive urgency (M [SD])	2.03 (0.59)	1.97 (0.55)	2.22 (0.54)	2.09 (0.54)	t(516)=-0.74	.459	t(516)=2.77	.006	t(516)=-0.80	.422
OCI-R (M [SD])	16.25 (12.18)	15.64 (12.40)	18.46 (12.36)	19.63 (13.75)	t(516)=0.80	.423	t(516)=1.40	.161	t(516)=-0.39	.696

Notes. Alcohol Use Disorder (AUD) criteria and diagnosis according to the Diagnostic and Statistical Manual of Mental Disorders (DSM)–5 [53] were assessed via a self-report questionnaire [54,55]. Drinking days, drinks per dinking day, and binge days were assessed via a Quantity Frequency Questionnaire according to the Munich Composite International Diagnostic Interview [56,57]. We removed implausible values from 'drinking days past 3 months' and 'drinks per drinking day in past 3 months' (> 7 drinking days / week; > 91.245 drinking days / 3 months; > 0 drinking days in past 3 months & 0 drinks per drinking day; drinks per weekday and weekend day > drinks per drinking day in past 3 months; > 50 drinks per drinking day in past 3 months) and from 'binge days in past 3 months' (> 7 binge days / week; > 91.245 binge days / 3 months). Smoking, cannabis, and other drug use were assessed using custom screening questions. If count values don't add up to 130 per column, the difference is made up by missing data. When no degrees of freedom for χ^2 tests are given, this indicates that *p*-values were simulated. AUDIT, Alcohol Use Disorder Identification Test [28,29]; CAS-A, Craving Automated Scale for Alcohol [58]; OCI-R, Obsessive Compulsive Inventory-Revised [61,62]; UPPS-P, Urgency-Premeditation-Perseverance-Sensation Seeking-Positive Urgency scale [59,60]. [a] Based on *n*=122 with valid data; [b] Based on *n*=129 with valid data; [c] Based on n=128 with valid data; [d] Based on *n*=124 with valid data; [e] Based on n=107 with valid data; [f] Based on n=127 with valid data; [g] Based on n=71 with valid data