# Learning Mechanisms to Predispose Risky Alcohol Drinking Behaviors During Young Adulthood

## DISSERTATION

for the degree of

**Doctor rerum naturalium**

**(Dr. rer. nat.)**

submitted to the

School of Science

Technischen Universität Dresden

by

**M. Sc., Hao Chen (陈灏)**

submitted in August, 2022

The dissertation was prepared from May, 2017 to August, 2022 under the supervision of Prof. Dr. med. Michael N. Smolka at the Section of Systems Neuroscience, Department of Psychiatry and Psychotherapy, Technischen Universität Dresden.

*For Grandma and Grandpa*

# Foreword

This thesis is based on publications which have already been published or manuscripts which are ready for submission to peer-reviewed journals. These publications have not been used in other theses and will not be used in future.

**Chapter 2 (Study 1)**

**Chen, H.**, Mojtahedzadeh, N., Belanger, M. J., Nebe, S., Kuitunen-Paul, S., Sebold, M., Garbusow, M., Huys, Q. J. M., Heinz, A., Rapp, M. A. & Smolka, M. N. (2021). Model-based and model-free control predicts alcohol consumption developmental trajectory in young adults: a 3-year prospective study. *Biological psychiatry*, *89*(10), 980-989.

**Contributions:** MNS, AH, QJMH and MAR were responsible for the study concept and design. SN, MS, MG, and SKP contributed to data acquisition. HC analyzed the data. HC, MNS and MAR interpreted the findings. HC drafted the manuscript. All authors critically reviewed content and approved the final version for publication.

**Chapter 3 (Study 2)**

**Chen, H**., Nebe, S., Mojtahedzadeh, N., Kuitunen-Paul, S., Garbusow, M., Schad, D. J., Rapp, M. A., Huys, Q. J. M., Heinz, A. & Smolka, M. N. (2021). Susceptibility to interference between Pavlovian and instrumental control is associated with early hazardous alcohol use. *Addiction biology*, *26*(4), e12983.

**Contributions:** MNS, AH, QJMH and MAR were responsible for the study concept and design. SN, MG and SKP contributed to data acquisition. HC, SN and MNS analyzed the data. HC, MNS and SN interpreted the findings. HC drafted the manuscript. MNS, SN, QJMH, SKP, MG, AH,

DJS and NM provided critical revision of the manuscript for important intellectual content. All authors critically reviewed content and approved the final version for publication.

**Chapter 4 (Study 3)**

**Chen, H.**, Belanger, M. J., Garbusow, M., Kuitunen-Paul, S., Huys, Q. J. M., Heinz, A., Rapp, M. A. & Smolka, M. N. (*major revision in Addiction Biology*). Susceptibility to interference between Pavlovian and instrumental control predisposes risky alcohol use developmental trajectory from ages 18 to 24

**Contributions:** MNS, AH, QJMH and MAR were responsible for the study concept and design. MG, SKP and HC contributed to data acquisition. HC analyzed the data. HC, MNS and MAR interpreted the findings. HC drafted the manuscript. All authors critically reviewed content and approved the final version for publication.

# Table of Contents

# List of Figures

# List of Tables

# List of Abbreviations

| | |
|---|---|
| AAL | Automated Anatomical Labelling |
| ADS | Alcohol Dependence Scale |
| AEQ | Alcohol Expectancy Questionnaire |
| ANOVA | Analysis of Variance |
| AUD | Alcohol Use Disorder |
| AUDIT | Alcohol Use Disorder Identification Test |
| AUDIT-C | Alcohol use disorders Identification Test Consumption Score |
| A-O | Action-outcome Association |
| BART | Balloon Analog Risk Task |
| BIC | Bayesian Information Criterion |
| BIS | Barrat Impulsiveness Scale |
| BMA | Bayesian Model Averaging |
| BMS | Bayesian Model Selection |
| BOLD | Blood-oxygen-level-dependent |
| BSPS | Blatant and Subtle Prejudice Scale |
| CFI | Comparative Fit Index |
| CS(s) | Conditioned Stimulus (stimuli) |
| CTQ | Childhood Trauma Questionnaire |
| dmPFC | Dorsomedial Prefrontal Cortex |
| DCM | Dynamic Causal Modeling |
| DLPFC | Dorsolateral Prefrontal Cortex |
| DMQ | Drinking motive questionnaire |
| DSbw | Digit Span Backwards Test |
| DSM | Diagnostic and Statistical Manual of Mental Disorders |

| | |
|---|---|
| DSST | Digit Symbol Substitution Task |
| EEG | Electroencephalogram |
| EPI | Echo-planar Imaging |
| ER | Error Rate |
| fMRI | Functional Magnetic Resonance Imaging |
| FRN | Feedback-related Negativity |
| FTND | Fagerström Nicotine Dependence Scale |
| FTQ | Family Tree Questionnaire |
| GLM | General Linear Model |
| IFG | Inferior Frontal Gyrus |
| lPFC | Lateral Prefrontal Cortex |
| LGCM | Latent Growth Curve Modeling |
| MB | Model-based |
| MF | Model-free |
| MNI | Montreal Neurological Institute |
| MPRAGE | Magnetization-Prepared Rapid Gradient-Echo |
| MRI | Magnetic Resonance Imaging |
| MWT | Mehrfachwahl-Wortschatz-Intelligenztest |
| M-CIDI | Munich Composite International Diagnostic Interview |
| OCDS | Obsessive Compulsive Drinking Scale |
| PIT | Pavlovian-to-instrumental Transfer |
| RFX | Random-effect Analysis |
| RL | Reinforcement Learning |
| RMSEA | Root Mean Square Error of Approximation |
| ROI | Region(s) of Interest |
| RPE | Reward Prediction Error |

| | |
|---|---|
| SAPAS | Standardised Assessment of Personality - Abbreviated Scale |
| SARSA | State-action-reward-state-action |
| SD | Standard Deviation |
| SMA | Supplementary Motor Area |
| SPF | Saarbrücker Persönlichkeitsfragebogen zur Messung von Empathie |
| SPM | Statistical Parametric Mapping |
| SRMR | Standardized Root Mean Square Residual |
| SRRS | Social Readjustment Rating Scale |
| SURPS | Substance Use Risk Profile Scale |
| S-O | Stimulus-outcome Association |
| S-R | Stimulus-response Association |
| TAS | Toronto Alexithymia Scale |
| TD | Temporal Difference |
| TMT | Trial Making Test |
| UCS(s) | Unconditioned Stimulus (stimuli) |
| vmPFC | Ventromedial Prefrontal Cortex |
| VLPFC | Ventrolateral prefrontal cortex |
| VS | Ventral Striatum |
| VTA | Ventral Tegmental Area |

## Abstract

Alcohol use disorder (AUD) is a mental disorder that negatively affects personal health and burdens the global health system. Alcohol-attributed harms can also extend beyond the drinkers to other people in the society through increased road traffic accidents and more interpersonal violent behaviors. The effects of this disorder make it crucial to investigate predisposing mechanisms in order to identify at-risk individuals and further develop novel interventions. Although aberrant learning and dysfunctions in decision-making have been observed in individuals with AUD, it is not yet clear whether they predispose the development of risky drinking behaviors or result from repetitive alcohol use. To disentangle this, we studied the drinking behaviors of a community sample comprising participants who were 18–24, which is when the prevalence of alcohol use typically peaks. This thesis investigates whether two types of learning mechanisms—the balance between goal-directed and habitual control and the susceptibility to interference between Pavlovian cues and instrumental behaviors—are associated with the development of risky alcohol drinking behaviors.

For Study 1, we assessed how goal-directed and habitual controls at 18 predispose alcohol use development over the course of 3 years. Goal-directed and habitual control, which are informed by model-based (MB) and model-free (MF) learning, were assessed with a two-step sequential decision-making task during functional magnetic resonance imaging. Three-year drinking trajectories were constructed based on the Alcohol Use Disorders Identification Test (AUDIT-C; assessed every 6 months) and a gram/drinking occasion measure (binge drinking score; assessed yearly). Latent growth curve models were applied to examine how the MB and MF controls were associated with the drinking trajectories. We found that MB control was negatively associated with the development of the binge drinking score trajectory. In contrast, MF reward prediction signals in the ventromedial prefrontal cortex and the ventral striatum (VS) were associated with a higher starting point and a steeper increase/less decrease in AUDIT-C, respectively.

For Study 2, we investigated the cross-sectional association between the susceptibility to interference between Pavlovian cues and instrumental behaviors and risky (binge) drinking behaviors at age 18. During a Pavlovian-to-instrumental transfer (PIT) task, the participants were instructed to “collect good shells” and “leave bad shells” while the appetitive (monetary

gain) or aversive (monetary loss) Pavlovian cues were presented in the background. The behavioral interference PIT effect was characterized by an increased error rate (ER) during incongruent trials ("collecting good shells" in the presence of an aversive Pavlovian cue or "leaving bad shells" during the presentation of an appetitive Pavlovian cue) in comparison to congruent ones. Overall, the individuals demonstrated a substantial behavioral PIT effect. Neural PIT correlates were found in the VS, dorsomedial, and lateral prefrontal cortices (dmPFC and lPFC, respectively). High-risk drinkers, in comparison to low-risk drinkers, exhibited a stronger behavioral PIT effect, decreased lPFC responses, and increased trend-level VS responses. Moreover, the effective connectivity from the VS to the lPFC during the incongruent trials was weaker for the high-risk drinkers, which indicates that the altered interplay between bottom-up and top-down neural responses may contribute to the poor interference control performance of this group.

During Study 3, we further examined whether the susceptibility to Pavlovian cues during conflict trials was associated with the development of drinking behaviors over 6 years from ages 18 to 24. The drinking behaviors were again constructed based on the AUDIT-C and the binge drinking score. The PIT task was assessed at ages 18 and 21. Following Study 2, the increased ER in the incongruent condition compared with the congruent condition (along with the neural responses in the VS, lPFC, and dmPFC during the incongruent trials) were included in the latent growth curve models as predictors. A stronger VS response during a conflict at age 18 was associated with a higher starting point in both drinking trajectories but was negatively associated with the development of the binge drinking score trajectory. At age 21, high ER and enhanced neural responses in the dmPFC were associated with a risky AUDIT-C trajectory that started to emerge and develop until age 24. Through exploratory cluster analyses of the drinking trajectories, we identified two subgroups: the drinking behavior in the "late riser" group escalated after age 21, whereas the drinking of "early peakers" culminated at this age and then declined. The late risers displayed enhanced dmPFC responses and higher ER during conflict at age 21. Interestingly, this group also exhibited an increased ER from ages 18 to 21.

Taken altogether, the unbalanced goal-directed to habitual control, informed by less MB and more MF control, appears to be a strong predisposing candidate mechanism that underlies the development of risky drinking behaviors during young adulthood. At age 18, the

susceptibility to interference between Pavlovian cues and instrumental behaviors was associated with risky drinking behavior. The development of risky drinking behaviors over the 6 years was associated with the behavioral interference PIT effect at age 21 and its change from ages 18 to 21. Researchers could further explore the dynamics in PIT to predict risky drinking behaviors in the future.

# Chapter 1: General Introduction

## 1.1 The Burden of Alcohol Use Disorder

Alcohol use disorder (AUD) is a mental disorder that negatively affects personal health and the global health system. According to a recent survey carried out in 27 countries or country regions between 2001 and 2015 (Glantz et al., 2020), the prevalence of lifetime alcohol use was 80% on average. Among these people, 10.7% developed AUD during their lifetimes, and 43.9% of these individuals had at least one other mental health disorder during their lifetimes. Moreover, the rates of suicidal behavior are three times higher in individuals with AUD than those without (Conner & Bagge, 2019).

Alcohol use can cause many chronic diseases and conditions other than AUD, and alcohol can also contribute to certain cancers and many cardiovascular diseases (Shield et al., 2014). Notably, the total volume of alcohol consumption plays a causal role in the onset or even culmination of these diseases (Rehm et al., 2017; Shield et al., 2014; Stevens et al., 2020). On the global scale, 3 million deaths were attributable to alcohol in 2016, which comprised 5.3% of deaths that year (Shield et al., 2020). Researchers have estimated that the average alcohol-attributable mortality after adjusting for age in 26 European countries is 10.1% for men and 3.3% for women (Janssen et al., 2020). As for the burdens that alcohol has been causing for the health system, some researchers demonstrated that alcohol-attributable hospitalizations in Canada during 2017 were even higher than the hospitalizations reported during the first 5 months of the COVID-19 pandemic (Stockwell et al., 2021).

Moreover, as indicated by a 26-year longitudinal study conducted in Finland, individuals who constantly experience high-frequency heavy drinking also experience more socioeconomic difficulties at age 42 (Berg et al., 2013). Other researchers also found during a Swedish longitudinal cohort that AUD casually influenced the receipt of social assistance and early retirement, as well as unemployment (Kendler et al., 2017). Critically, the harm caused by alcohol is not only restricted to the drinkers, but also negatively affect other people. For example, alcohol intoxication can lead to more road traffic accidents and more violent behaviors (Kraus et al., 2019).

Given the high risks of AUD's comorbidity with other mental health disorders, its high mortality rate, the burdens that AUD has been causing on the healthcare and social systems, along with the harms to other people, the attempt to identify mechanisms that predispose risky drinking behaviors or AUD at an early stage is vital.

## 1.2 Identifying Individual Vulnerabilities

Like other addictive disorders, AUD can be regarded as a disorder that follows specific developmental processes (Miller, 2018). From a developmental perspective, individuals tend to start experimenting with alcohol during adolescence, and the prevalence of alcohol use reaches its peak between ages 18 and 24 (Miller, 2018). Therefore, it is possible to identify common risky drinking patterns. From the neurocircuitry perspective, the typical development cycle follows three stages: binge/intoxication, withdrawal/negative affect, and preoccupation/anticipation (Koob & Le Moal, 2005; Koob & Volkow, 2010, 2016). Accordingly, the development of addiction involves a shift from the positive reinforcing effect of drugs to the "dark side", in which an individual's drug-seeking behavior is mainly driven by the motivation to reduce aversive feelings that occur during withdrawal. To identify individual vulnerabilities and develop preventions, one should study when risky drinking behaviors develop prominently, determine how to assess these risky drinking patterns, and define the mechanisms that contribute to this development. The following sections aim to evaluate and provide a general framework for tackling these questions.

## 1.3 When? Early Young Adulthood Is the Key

Emerging adulthood, typically ages 18–25 in industrialized and postindustrial countries, is a period when individuals experience heightened identity exploration, changes in subjective perceptions, and instabilities and also face many possibilities (Arnett, 2000). Due to the distinctive features of this period in life, the prevalence of substance use during this period is also at its highest (Arnett, 2005; Miller, 2018; Sussman & Arnett, 2014).

On the one hand, the contextual change that occurs when people attend college and fulfill their adult roles may largely contribute to the emergence of heavy drinking and drinking-related problems (Schulenberg & Maggs, 2002). On the other hand, the brain continues to develop through late adolescence and further into early young adulthood, with higher-order cortices developing later than the sensorimotor cortices (Giedd et al., 1999; Gogtay et al.,

2004); cognitive control also continues to evolve from adolescence to the early 20s (Shulman et al., 2016). The neurological development that takes place around these ages makes people vulnerable to alcohol exposure during the transition period from adolescence to young adulthood (Brown et al., 2008). While risky alcohol use is a common problem among young adults (Schulenberg & Maggs, 2002), heterogeneous, distinctive binges and heavy drinking patterns have also been observed during the young adulthood period (Jackson et al., 2008; Tucker et al., 2003; Windle et al., 2005).

Overall, I believe that young adulthood is a period that can introduce more dynamic and distinctive drinking patterns. It is thus important to understand why hazardous alcohol use escalates and why AUD becomes chronic and newly emerges for certain people. In contrast, alcohol-related problems are remiss without any treatment for others. To address these questions, we adopted a longitudinal design of over 6 years that includes participants from ages 18 to 24 to assess different drinking behaviors every 6 months or yearly to capture the dynamics during this critical period.

## 1.4 How? Assess the Intermediate States Towards AUD

After identifying the critical period in life that is important for identifying individual vulnerabilities, one also needs to assess the at-risk statuses of individuals. Our approach to capturing the development during this stage was to model the developmental trajectories of a few drinking variables as proxies or intermediate states that could lead to developing AUD. This section focuses on two variables—binge drinking and alcohol consumption scores as assessed by AUDIT (AUDIT-C)—and demonstrates why they could measure hazardous drinking behavior.

### 1.4.1 Binge Drinking

According to the World Health Organization, a cut-off of 60 g of ethanol intake per drinking occasion can be used to define binge drinking (Stockwell et al., 2000; World Health Organization, 2019). Binge drinking is regarded as a typical drinking pattern during adolescence and young adulthood in Western countries, with a sharp increase in prevalence from adolescence to young adulthood (Jones et al., 2018; Lees et al., 2019).

On the cross-sectional level, deficits in decision-making and inhibition are significant factors that are associated with binge drinking, according to researchers who conducted a large-scale meta-analysis to investigate young binge drinkers (1,313 binge drinkers, mean age = 18.83) (Lees et al., 2019). This meta-analysis thus demonstrated the critical links between neurocognitive deficits and binge drinking during young adulthood. From the longitudinal perspective, smaller brain volume in the dorsolateral prefrontal cortex (DLPFC) during the early adolescence period (ages 12 to 14) was found to predict more binge drinking during the follow-up after an average of 1.7 years (ranging from 1 to 7 years) (Brumback et al., 2016). Additionally, weaker brain responses during the working memory and inhibitory control tasks in the frontoparietal regions during the early adolescence period could also predict binge drinking behaviors and the initiation of binge drinking, respectively, during late adolescence or early young adulthood (Squeglia et al., 2012; Wetherill et al., 2013).

The evidence thus suggests that binge drinking behavior during young adulthood is associated with poorer executive functions that heavily rely on the DLPFC functions, such as inhibitory control. More importantly, given the association between binge drinking behavior and neurobiological changes, the development of binge drinking behavior has been suggested to be a crucial intermediate state that can lead to the development of AUD later in life (Cservenka & Brumback, 2017; Jones et al., 2018).

### 1.4.2 General Alcohol Consumption (AUDIT-C)

The AUDIT-C, which assesses the quantity and frequency of drinking and binge drinking, has also been suggested to be an effective tool for identifying AUD and risky drinking behavior (Dawson, 2011; Dawson et al., 2005). In a study that featured 5,401 university students (ages 17–25), an AUDIT-C score cut-off of 7 for females and 8 for males was found to be optimal for identifying hazardous drinking among college students (Verhoog et al., 2020). Thus, the AUDIT-C seems to be another well-qualified candidate for characterizing risky-drinking behavior during young adulthood.

So far, only a few researchers have attempted to identify the associated neurobiological factors with the AUDIT-C during young adulthood. In one electroencephalogram (EEG) study, researchers found that the attenuated feedback-related negativity (FRN) and enhanced feedback-locked P3 components, which indicate reward prediction error (RPE) signals after

receiving rewards, were associated with higher alcohol consumption scores (Cao et al., 2021). Researchers who conducted another EEG study assessed alcohol consumption with a daily drinking questionnaire that integrated the frequency and quantity during a typical (binge) drinking occasion. They found a negative association between FRN and alcohol consumption (Soder et al., 2019). Interestingly, with the same task (Balloon Analogue Risk Task) as in Soder et al. (2019), the differences in the feedback-related FRN and P3 components were not found between binge and non-binge drinkers (Lannoy et al., 2017). These studies thus indicate that RPE processing is associated with alcohol consumption, and this association cannot be seen by solely assessing binge drinking behavior. Therefore, it is reasonable to characterize the developmental trajectories of the consumption score, in addition to the binge drinking trajectory, as another intermediate state that leads to developing AUD.

## 1.5 What? General Theoretical Framework

Finally, one must ask the "what" question to define the mechanisms of interest to identify individual vulnerabilities. This is a more challenging question since the development of alcohol addiction involves many mechanisms, and many theories have been proposed to explain the processes (see Bickel et al. [2018] for a review of neurobehavioral theories; Redish et al. [2008] for a unified decision-making framework). Our goal was to look for mechanisms associated with risky drinking patterns during young adulthood, which is when drinking behaviors start to accelerate and risky drinking patterns begin to emerge. During this initial binge/intoxication phase, alcohol produces a reinforcing effect by prominently increasing the dopamine release in the basal ganglia (Urban et al., 2010). Individuals thus need to maintain proper baseline dopamine-related functions during this phase. Notably, the dopaminergic system can encode a broad range of information that involves learning and decision-making, such as approach and avoidance behaviors and stimulus-response associations (Cohen et al., 2012; Schultz, 2007a; Schultz et al., 2000). For the "what" question, therefore, I focused on dopamine-related theories (as summarized by Bickel et al. [2018]), which emphasize the neurobehavioral processes that are related to dopamine signaling.

Dopamine-related theories are rooted in a long-standing learning history that dates back to the famous Pavlovian conditioning experiments (Pavlov, 1960) that demonstrated how organisms can learn from past experiences. After Pavlov's experiments, researchers indicated

that different types of drugs share common biological mechanisms despite their different molecular targets, with the dopaminergic system being the best candidate given its role in positive reinforcement (Wise, 1987, 1988). Later experiments conducted by Shultz et al. demonstrated the critical role of midbrain dopaminergic neurons in computing RPE that facilitates learning (Schultz, 1986; Schultz et al., 1993; Schultz et al., 1997). These early works have convincingly demonstrated how learning, the dopamine system, and addiction might be linked together; therefore, learning theories have played significant roles in addiction literature (Everitt & Robbins, 2016).

## 1.6 Goal-Directed and Habitual Instrumental Control

In the formal associative learning framework that was developed by Anthony Dickinson (Dickinson & Balleine, 1994), the instrumental process associates actions with outcomes (A-O), whereas the Pavlovian system learns the association between contextual stimuli and outcomes (S-O). Instrumental behavior is regarded as goal-directed when the outcome is represented as a goal for the agent. Robbins and Everitt (1999) later incorporated the concept of addiction into their associative learning framework by suggesting that drug addiction is a form of aberrant learning. Moreover, the authors suggested that the instrumental process also requires the learner to learn the stimulus-response (S-R) association in addition to the A-O association, where the response is not sensitive to outcome changes. The two forms of instrumental learning, A-O and S-R learning, correspond to goal-directed and habitual control, respectively.

Hogarth et al. (2012) reviewed and combined the previous learning theories and the empirical evidence into a general framework in the context of addiction (Figure 1). They summarized the experimental findings in animal and human studies and proposed a transitional model of addiction. Concerning instrumental learning, initial drug-seeking involves goal-directed control (A-O). After extensive training or extensive drug use in a real-world setting, an individual's behavioral response is mainly driven by habitual control; their S-R associations dominate their behavior and drive compulsive drug use at this stage.

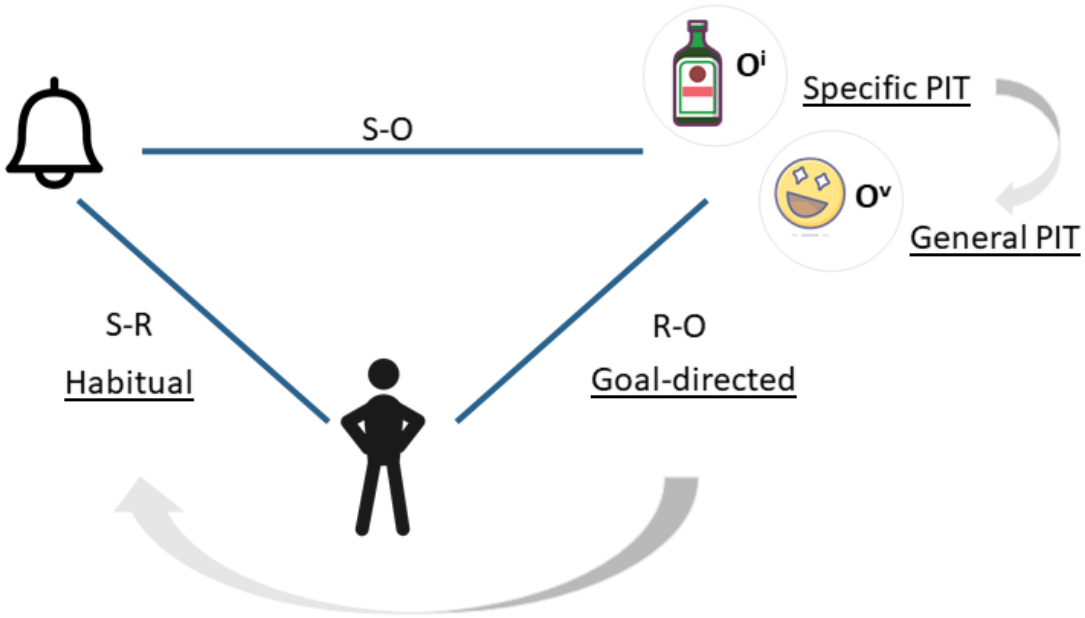


**Figure 1: The transition model during the development of addiction**. The development of addictive behavior involves a shift from goal-directed to habitual control. Moreover, drug stimuli initially influence instrumental behaviors through a specific Pavlovian-to-instrumental transfer (PIT) process, but after continuous drug exposure, drug cues could elicit actions through a general PIT process (adapted from Hogarth et al. [2012], Figure 1).

### 1.6.1 Brief Introduction to the Reinforcement Learning Framework

Intriguingly, the learning theory in psychology and neuroscience corresponds well with the reinforcement learning (RL) theory in the computational science field (see Sutton and Barto [1998] for more about RL theory). According to the RL theory, learning is essentially about selecting actions that maximize rewards or minimize punishment. The rewards that are achieved during reinforcement learning can be immediately available but could also be delayed, which requires the agent to calculate all future rewards. Accordingly, the agent selects actions based on trial-and-error or the expectation of delayed rewards. Neuromodulator dopamine, with the signaling of the RPE signal (Schultz, 2007b; Schultz et al., 1997), connects the RL theory to learning and decision-making processes in the brain. Notably, the RL theory can provide a normative framework for understanding behaviors since an optimal action that can be performed to achieve specific goals in a certain environment could be derived from the models (Niv, 2009).

To provide a basis for the RL model, I recommend starting with the temporal difference (TD) learning rules (Sutton, 1988). In TD learning, the state of the world ($s_t$) progresses with time step $t$, which represents a trial or smaller compartments within a trial; $r(s_t)$ represents the reward incurred at the state $s_t$. In addition to the immediately available rewards, the agent also considers all expected future rewards. All future rewards are discounted to the current state with the discounting rate $\gamma$. Defining the value in state $s_t$ as $V(s_t)$, the expectation of rewards, given the state $s_t$, could be written in the following way:

$$V(s_t) = r(s_t) + E[\gamma r(s_{t+1}) + \gamma^2 r(s_{t+2}) + \gamma^3 r(s_{t+3}) \dots | s_t]$$
$$= r(s_t) + \gamma E[r(s_{t+1})|s_t] + \gamma^2 E[r(s_{t+2})|s_t] + \gamma^3 E[r(s_{t+3})|s_t] \dots \quad (1)$$

Similarly, for the next state $s_{t+1}$:

$$V(s_{t+1}) = r(s_{t+1}) + \gamma E[r_{t+2}|s_{t+1}] + \gamma^2 E[r_{t+3}|s_{t+1}] + \cdots \quad (2)$$

By integrating Equation (2) into (1), one reaches what is known as the *Bellman equation* (Bellman, 1957):

$$V(s_t) = r(s_t) + \gamma E[V(s_{t+1})|s_t] \quad (3)$$

When the acquired reward is different from the expectation, the discrepancy between the two sides could be expressed as the temporal difference prediction error:

$$\delta_t = r(s_t) + \gamma E[V(s_{t+1})|s_t] - V(s_t) \quad (4)$$

Note that the $E[V(s_{t+1})|s_t]$ term in Equations (3) and (4) defines probability-weighted average expectations across all possible successor states; however, when the agent does not know the environment, they could also learn by sampling the rewards from one state to another. This represents model-free (MF) RL (Barto et al., 1989; Bertsekas & Tsitsiklis, 1996), where the update is done at every step through the trial-and-error prediction error. The reward prediction error could be written as an approximation of Equation (4):

$$\delta_t = r_t + \gamma V(S_{t+1}) - V(s_t) \quad (5)$$

TD learning thus offers a way to describe how an agent could learn to predict the optimal values of different events even without knowing the dynamics of an environment. As mentioned previously, the agent's goal is to select optimal actions that maximize their future rewards or minimize punishments. To achieve this, the agent needs to assign credits (Sutton

& Barto, 1998) to the actions that have led to the desired outcomes and repeat these actions accordingly in specific states. One commonly adopted method is to learn the values of the state-action pairs, which could be denoted as *Q(s, a)*. In Q-learning (Watkins, 1989), the TD prediction error could be written in the following way:

$$\delta_t = r_t + \max_a \gamma Q(s_{t+1}, a) - Q(s_t, a_t) \quad (6)$$

Here, instead of only learning the value of one state, the agent learns the values of state-action pairs. The value of state $s_{t+1}$ is considered to be the value associated with the best action taken in this successor state, even though the agent will not actually take this action. The Q-learning model is thus considered an "off-policy" method. In contrast, the agent could also adopt an "on-policy" state-action-reward-state-action (SARSA) method:

$$\delta_t = r_t + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t) \quad (7)$$

With the SARSA method, the agent calculates the TD error through the actual action to be taken instead of the action that gives the highest value in state $s_{t+1}$. In both Q-learning and SARSA, the update of the state-action pairs could be done by performing the following equation:

$$Q(s_t, a_t)_{new} = Q(s_t, a_t)_{old} + \alpha\delta_t \quad (8)$$

Importantly, α represents the learning rate parameter, which ranges between 0 and 1, and describes how quickly the update is done.

### 1.6.2 MB and MF Learning

So far, I have described trial-and-error learning and how to formulate it within the RL framework. This type of learning is typically regarded as MF learning, where the agent learns through cached values of state-action pairs. However, this method is computationally cheap since it is slow and inflexible. In contrast, instead of selecting actions based on trial-by-trial outcomes, one could also learn by building an internal model.

Consider an agent who tries to maximize their expected rewards by selecting actions in an uncertain environment. The value given by certain actions could be written as the following equation (Daw & O'Doherty, 2014):

$$Q(a) = \sum_s P(s|a)r(s) \quad (9)$$

In this equation, the selected actions result in different states with probabilities $P(s|a)$, and the reward in each state is *r(s)*. The values of specific actions could be learned through the trial-and-error approach based on the Q value on the left-hand side, which was previously described as the MF learning (no internal environment model). Conversely, an agent could also build an internal model based on the right-hand side of the equation; that is, they could plug in the probability (or likelihood) of entering certain states through actions and the outcome associated with the state that represents model-based (MB) RL. More formally, the MB system applies the "tree-search" method and updates the values of the state-action pairs through forward-planning. This type of learning is computationally expensive but may provide more accurate predictions when the environment is complicated and dynamic.

Given that both MF and MB learning can include pros and cons during action selection, Daw et al. (2005) proposed that the brain arbitrates between the two systems when an environment is uncertain. Depending on the certainty level, the recommendation from the system with higher certainty is acted upon. Daw et al. (2005) also explained how MB and MF learning function during the outcome devaluation manipulation (Dickinson, 1985), which typically identifies goal-directed behavior. During the outcome devaluation procedure, actions are first paired with outcomes through instrumental learning. Afterward, one of the outcomes is made less desirable (for example, through satiation). When relying on the MF system, the agent still repeats the action associated with this outcome since the value associated with such an action has been cached. However, when applying the tree-search method based on the MB system, the agent is sensitive to the devaluation when inserting the devalued *r(s)* into the right-hand side of the equation (9).

Daw et al. (2005) demonstrated that MF learning, which is defined in the RL framework, could correspond to habitual behavior, which is insensitive to the outcome devaluation, while MB learning is closely related to goal-directed behavior, which is sensitive to the change of outcomes. The MB and MF RL thus offer a normative framework to study goal-directed and habitual control. Following this notion, Daw et al. (2011) proposed a two-step task that approaches goal-directed and habitual control through the assessment of MB and MF learning during RL. The two-step task thus offers an excellent opportunity to study the association between risky alcohol use and unbalanced goal-directed and habitual behavior within the normative RL framework.

### 1.6.3 Unbalanced MB and MF Control with Alcohol Use: Empirical Evidence

MB and MF control, when assessed with the two-step task, were investigated in participants with AUD. In one study, Voon et al. (2015) found no differences between abstinent AUD patients (with abstinence periods ranging from 2 weeks to 1 year) and healthy controls concerning the strategies used in the two-step task. Conversely, recently detoxified AUD patients (3 weeks on average) exhibited less MB behavioral control than the control participants did (Sebold et al., 2014). However, this result was not replicated in the whole sample. Instead, relapsers were found to display fewer MB neural responses in the medial prefrontal cortex (Sebold et al., 2017). Regarding this, some researchers found that AUD patients demonstrated deficits in updating their alternative choice options when engaged in a reversal-learning task, which could be linked to the deficits in the MB control system (Reiter et al., 2016a). Taken together, the association between AUD and goal-directed and habitual control is mixed but seems to indicate that AUD patients tend to have less MB control.

In the non-clinical populations, less MB control was found to be associated with binge drinking (Doñamayor et al., 2018) and AUDIT scores in a large general population sample (Gillan et al., 2016). Reiter et al. (2016b) did not find an association between a family history of alcohol dependence with the strategies used in the two-step task. However, this could have been due to their small sample size (*N* = 20).

Overall, although there has been evidence regarding the association between alcohol consumption, binge drinking, and less MB control, it is not clear whether such an association works as a predisposing mechanism or evolves from repetitive alcohol use. Additionally, although the evidence regarding the association between MF control and alcohol use is limited, the positive association between MF control and alcohol use could be assumed on the theoretical level. Moreover, it is also plausible that the MF might initially be linked to risky alcohol use but is not maintained over the course of repetitive alcohol use.

## 1.7 Pavlovian-to-Instrumental Transfer (PIT)

### 1.7.1 PIT Mechanisms

In addition to instrumental learning, (conditioned) drug cues can elicit craving and drug-seeking behavior (Robinson & Berridge, 1993), which emphasizes the importance of the

Pavlovian system. Formally, the Pavlovian system learns the association between cues (conditioned stimuli [CSs]) and rewards or punishments as outcomes (unconditioned stimuli [USs]). The learned (S-O) association could elicit a conditioned response, even though the behavioral response has not been learned explicitly. In the context of alcohol use, when the street view of one's favorite bar has been paired with a few glasses of beer, several lively conversations, and joy and laughter, this environmental cue could elicit conditioned responses such as alcohol craving. Indeed, researchers who conducted previous cue-reactivity studies found that showing alcohol cues to participants with AUD could lead to higher self-reported cravings, stronger physiological responses such as an increased heart rate (Carter & Tiffany, 1999; Witteman et al., 2015), pupil dilation (Kvamme et al., 2019), a stronger approach tendency (Wiers et al., 2014), and more attention biases towards alcohol cues (Vollstädt-Klein et al., 2012). These studies thus demonstrate the influential roles of Pavlovian cues in eliciting conditioned responses to drug-seeking.

However, conditioned responses such as craving and attention biases are not enough to drive an individual to seek drugs in daily life. How do such cues influence an individual to accomplish related goals? Some researchers have suggested that Pavlovian cues could influence goal-directed instrumental behavior through a PIT process (Corbit et al., 2007; Holmes et al., 2010). In the example just described, the person who just passed by their favorite bar may initiate a sequence of goal-directed behaviors such as sending messages in a group chat to organize a drinking event, entering the bar, and ordering drinks. Two types of transfer, specific and general transfers, can be identified (Hogarth et al., 2012). During a specific transfer, after the contingency of the stimulus and the outcome (identity) are identified, this $S\text{-}O^{i}$ association can influence the instrumental behavior that is associated with the same outcome. For example, pictures of beer can drive people to purchase beer at the supermarket. In contrast, during a general transfer, the incentive value of the outcome is learned; the learned $S\text{-}O^{v}$ association could enhance an individual's motivational state, which also influences their other available instrumental behaviors, even though these behaviors were not paired with the same outcome previously. For example, people may consider taking out a bottle of wine after preparing a decent meal.

### 1.7.2 Theories to Account for General and Specific PIT Effects

When it comes to the theories that explain how Pavlovian cues influence ongoing instrumental behavior, the incentive sensitization theory (Robinson & Berridge, 1993) is a long-standing theory that has been used to describe the general PIT processes, mainly in animal studies. This theory suggests, with the established association between CSs and USs, that when encountering CSs, the mesocorticolimbic circuits could generate incentive salience signals based on an individual's current neurological and physiological states so that the CSs and the USs become more "wanted" (Dayan & Berridge, 2014; Flagel et al., 2011; Mahler & Berridge, 2012; Robinson & Berridge, 2013; Robinson & Berridge, 1993). This incentive salience signal that acts as a response to Pavlovian cues could also boost an individual's motivation to elicit motivational responses. Incentive salience can thus trigger instrumental actions to obtain rewards or avoid threats to account for the PIT mechanism (Colwill & Rescorla, 1988; Holland, 2004; Mahler & Berridge, 2012; Peciña & Berridge, 2013).

When evaluating this theory in the context of drug use, one should note that drug stimuli could enhance the mesolimbic dopamine signaling and thus increase the preference for drug-seeking by enhancing the incentive salience of such events. After repetitive drug use, individuals may dissociate "wanting" (the incentive salience) and "liking" (the subjective euphoric experience). Consequently, they may continue to take drugs due to the amplified "wanting," even though the subjective euphoric experience of such behavior has become diminished (Berridge & Robinson, 2016; Robinson & Berridge, 1993, 2008).

Conversely, the motivational or arousing effects of drug cues could not explain the specific PIT effect in which the outcome (identity) also needs to be encoded. There is still ongoing debate about which theories could best account for the available experimental data in the field (see Mahlberg et al. [2019] for a detailed review of the theories). The critical debate concerns whether instrumental responses are sensitive to the outcome of devaluation procedures (i.e., the decreased response rate that is a result of the devaluation). This debate has primarily been caused by mixed experimental findings: Mahlberg et al. (2021) assessed the previously published studies and found eight studies that suggest that PIT effects are not sensitive to the devaluation procedure (as examples: Hogarth, 2012; Hogarth & Chase, 2011; van Steenbergen et al., 2017), while eight stated the opposite (refer to: Seabrooke et al., 2019; Seabrooke et al., 2017). When it is assumed that the PIT effects are insensitive to the

devaluation, the S-O-R theory should be considered (Balleine & O'Doherty, 2010; Balleine & Ostlund, 2007). This theory suggests that the Pavlovian system learns the S-O association, while the instrumental learns bidirectional associations between responses and outcomes (R-O/O-R). Therefore, the outcome representation is activated through the S-O association when a stimulus is presented, which triggers the response through the R-O/O-R association. Since this theory suggests that a Pavlovian stimulus only activates the sensory properties of the outcome, which then triggers the instrumental responses through the S-O-R chain, the elicited response is then insensitive to the devaluation procedure.

More recently, one caveat in the experimental design has been pointed out: instrumental learning can be biased if participants prefer one outcome over the other since this can lead to a biased baseline of instrumental responding rates. In this case, the change in the response rate after outcome devaluation is also biased. Researchers who corrected this baseline bias found that the PIT effects are not sensitive to the devaluation procedure (Seabrooke et al., 2019; Seabrooke et al., 2017). Accordingly, propositional theory (Mahlberg et al., 2021; Seabrooke et al., 2019; Seabrooke et al., 2016; Seabrooke et al., 2017; Seabrooke et al., 2018) has been proposed to account for these recent findings. The propositional theory suggests that both the S-O association (Pavlovian) and the R-O association (instrumental) are learned explicitly. During a transfer test, participants infer which outcomes are available according to the Pavlovian cues that are presented and which instrumental response can be used to acquire the outcomes. Therefore, this theory indicates that specific PIT effects represent a goal-directed, controlled process rather than an automatic one.

Overall, evidence is accumulating in the direction that the outcome-specific PIT could be regarded as a goal-directed process. In contrast, a strong general PIT effect (i.e., a higher motivational effect that is elicited by Pavlovian cues) may enhance behavioral sequences towards alcohol-seeking—the S-R link of habitual alcohol-seeking behavior is then formed after multiple repetitions. Therefore, once the drug cues acquire the general motivating effect, they could enhance the habitual drug-seeking behaviors, which would make related habits more persistent and compulsive (Everitt & Robbins, 2016). Hogarth et al. (2012) also proposed that the development of compulsive alcohol use may also involve a shift specific to general PIT processes, in addition to the shift from goal-directed to habitual control, as mentioned previously. Specifically, during early training, the drug cues produce specific transfer effects;

after an individual undergoes extensive drug exposure, they may not associate the cues with the specific identity of the drugs but rather with the value or the rewarding effect of the drugs. Nevertheless, these theories suggest that general and specific PIT processes play essential roles in developing addictive behaviors. The association between general PIT and compulsive alcohol use may be more substantial given its tighter link to habitual behaviors.

### 1.7.3 Single-Lever PIT and AUD: Empirical Evidence

The associative learning theories provide a framework that incorporates both general and specific forms of PIT by emphasizing the roles of appetitive cues on approach behaviors. As one type of the PIT paradigm, participants in one study needed to select from two approach behaviors: press one button for food and the other button for drugs during the presentation of Pavlovian cues that were paired with drug or food rewards previously (Hogarth et al., 2007). In contrast, a "single-lever" PIT paradigm with monetary feedback has been proposed to evaluate the influence of Pavlovian cues on instrumental behaviors as a more general mechanism (Garbusow et al., 2014). In a study that utilized this paradigm, to gain more monetary reward, the participants needed to learn to press one button to "collect good shells" and withhold responses to "leave bad shells." Pavlovian cues were appetitive (monetary rewards), aversive (monetary losses), or neutral. During the transfer, the participants needed to "collect good shells" and "leave bad shells" during the presentation of the different types of Pavlovian cues. So far, it is not clear whether the PIT effect in this PIT task reflects general or specific PIT. However, according to evidence from animal studies, it is reasonable to consider it as a general PIT effect (Cartoni et al., 2016).

Concerning alcohol use, the valence of Pavlovian cues had a more substantial impact on instrumental behavior (more button presses) for participants with AUD in comparison to the matched control group (Garbusow et al., 2016; Garbusow et al., 2014; Schad et al., 2019), as well as for high-risk versus low-risk drinkers (Garbusow et al., 2019).

### 1.7.4 An Alternative View of PIT: the Interference Control Perspective

The interference control perspective considers the selection between instrumental approach and avoidance behaviors and the influence of both appetitive and aversive Pavlovian cues. It is worth considering how the instrumental and Pavlovian systems interact. Huys et al. (2011) demonstrated that Pavlovian stimuli selectively motivate instrumental behavior according to

the intrinsic valences of the behavior: appetitive Pavlovian cues promote approach instrumental behaviors and inhibit avoidance behaviors, and the reverse is true for aversive Pavlovian cues. Given this interaction, investigating the conditions in which Pavlovian cues are not concordant to the required instrumental behaviors should be of particular interest.

The school of literature that assesses learning during a so-called valenced go-no-go task (Guitart-Masip et al., 2012) may offer some insights regarding the interactions between instrumental and Pavlovian controls, especially during a conflict. In this task, Pavlovian cues are embedded in the trial-and-error reinforcement learning process. Four different cues create four experimental conditions: go-to-win, go-to-avoid-losing, no-go-to-win, and no-go-to-avoid-losing conditions. Per definition, the Pavlovian system tends to approach when there winning is possible but avoids considering potential losses. The instrumental system, on the other hand, fully learns cue-action-outcome associations. Therefore, conflict could be elicited in the go-to-avoid-losing condition: the instrumental system knows the "go" action would be the correct response, but the Pavlovian control would suggest the "no-go" action since there is potential losing. This conflict also happens in the no-go-to-win condition. Some researchers have found that the accuracy in the conflict conditions is lower than in the other two non-conflict conditions (Ereira et al., 2021; Guitart-Masip et al., 2012).

Notably, in one EEG study, the researchers found that the ability to overcome the Pavlovian bias when it conflicted with the required instrumental behavior was also associated with the fontal theta power, which is sensitive to conflict (Cavanagh et al., 2013). Moreover, the motor and lateral prefrontal areas were found to synchronize with the midfrontal regions to reduce the Pavlovian bias, along with the midfrontal theta power (Swart et al., 2018). Taken together, although this task assesses the interaction between the two types of controls during the learning process, it provides compelling support for the notion that interference control is needed when Pavlovian cues conflict with ongoing instrumental behavior. The involvement of the lateral and midfrontal areas also supports the view that Pavlovian and goal-directed instrumental behavior may interact via cognitive control (Yee & Braver, 2018).

Although PIT tasks typically assess how Pavlovian cues influence instrumental behaviors after both Pavlovian and instrumental learning is achieved, the theories proposed for the valenced go-no-go task could potentially be extended to the PIT task. Formally, by adapting the action-valence axes proposed by Guitart-Masip et al. (2014) (Figure 2), one can understand how

instrumental control learns approach or avoidance behaviors during instrumental training (along the y-axis). In contrast, Pavlovian cues elicit conditioned responses according to the associated outcomes; they elicit approach tendencies (green line) when a cue is paired with a reward and avoidance tendencies when experiencing loss is possible (red line). Therefore, respectively, appetitive and aversive Pavlovian cues can interfere with instrumental avoidance and approach behaviors.

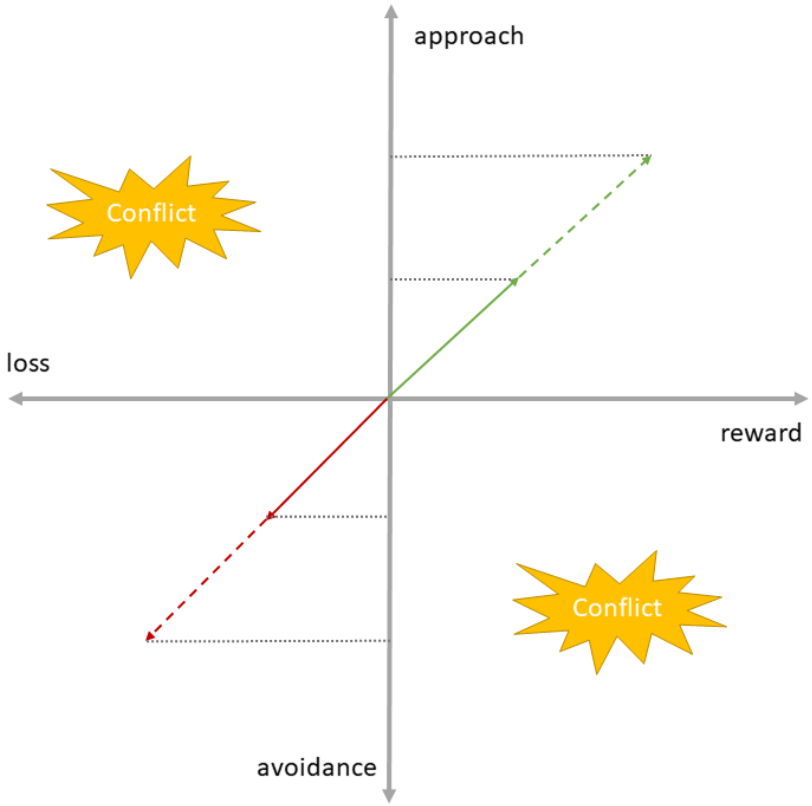


**Figure 2: Interaction between Pavlovian and instrumental control** (adapted from Guitart-Masip et al. [2012], Figure 1). Green line represents approach tendencies elicited by Pavlovian cues, while red line represents avoidance tendencies. The dashed lines indicate stronger influence of Pavlovian cues.

Being unable to perform instrumental behaviors adaptively could also be an important factor that is associated with AUD. Indeed, when analyzing a PIT task from an interference control perspective, Sommer et al. (2017) found that the AUD participants made more errors when they needed to inhibit their instrumental behaviors during the presentation of appetitive Pavlovian cues. Consistent with this finding, future relapsers failed to correctly inhibit their instrumental actions during the presentation of the appetitive Pavlovian cues (Sommer et al.,

2020). Thus, further investigating whether this interference effect of Pavlovian cues could work as a predisposing mechanism towards compulsive alcohol use is promising.

## 1.8 Research Questions

This thesis aims to investigate whether the unbalanced goal-directed and habitual control, the stronger susceptibility to the interference between Pavlovian and instrumental controls, and the associated neural mechanisms predispose risky drinking development across young adulthood. Risky drinking behaviors as assessed by AUDIT-C (every half-a-year started at age 18.5) and binge drinking scores (assessed yearly) were characterized using the latent growth curve models. Upward trends (i.e., increasing slopes) in these trajectories can be regarded as intermediate hazardous states that can develop into compulsive alcohol use.

For Study 1, we examined whether MB (goal-directed) and MF (habitual) learning with the two-step task at age 18 could be used to predict the participants' drinking trajectories from ages 18 to 21 with latent growth curve models. We assumed that less MB control and more MF control, as indicated by behavioral scores and neural RPE signals, would be associated with riskier alcohol use.

During Study 2, we aimed to establish the relationship between interference PIT effects and binge drinking behavior cross-sectionally at age 18. We also aimed to uncover the underlying neural mechanisms of interference control that are elicited by the conflict between Pavlovian cues and ongoing instrumental behaviors with functional magnetic resonance imaging (fMRI). We hypothesized that the conflict would elicit neural responses in brain regions that are associated with cognitive control, such as the lateral and dorsomedial prefrontal cortices (lPFC and dmPFC), along with the ventral striatum (VS) and amygdala responses, which have been observed in previous PIT studies. We also expected high-risk drinkers (binge drinkers) to make more erroneous responses (higher ER) during the incongruent trials, along with exhibiting stronger neural responses in the VS and weaker lPFC and dmPFC responses during interference processing.

For Study 3, we assessed whether the interference effect of the Pavlovian cues on both behavioral and neural levels could predispose the hazardous drinking trajectory over the 6-year follow-up from ages 18–24. Two PIT assessments at ages 18 and 21 were conducted for the prediction. Specifically, we hypothesized that the increased ER, along with higher neural

responses in the VS and lower responses in the lPFC and dmPFC, would predispose riskier drinking trajectories over the 6 years.

# Chapter 2: Goal-Directed and Habitual Control with the Three-Year Drinking Trajectory (Study 1)

This chapter has been published as the following:

**Chen, H.**, Mojtahedzadeh, N., Belanger, M. J., Nebe, S., Kuitunen-Paul, S., Sebold, M., Garbusow, M., Huys, Q. J. M., Heinz, A., Rapp, M. A. & Smolka, M. N. (2021). Model-based and model-free control predicts alcohol consumption developmental trajectory in young adults: a 3-year prospective study. *Biological psychiatry*, *89*(10), 980-989.

## 2.1 Abstract

**Background:** A shift from goal-directed toward habitual control has been associated with alcohol dependence. Whether such a shift predisposes to risky drinking is not yet clear. We investigated how goal-directed and habitual control at age 18 predict alcohol use trajectories over the course of 3 years.

**Methods:** Goal-directed and habitual control, as informed by model-based (MB) and model-free (MF) learning, were assessed with a two-step sequential decision-making task during functional magnetic resonance imaging in 146 healthy 18-year-old men. Three-year alcohol use developmental trajectories were based on either a consumption score from the self-reported Alcohol Use Disorders Identification Test (assessed every 6 months) or an interview-based binge drinking score (grams of alcohol/occasion; assessed every year). We applied a latent growth curve model to examine how MB and MF control predicted the drinking trajectory.

**Results:** Drinking behavior was best characterized by a linear trajectory. MB behavioral control was negatively associated with the development of the binge drinking score; MF reward prediction error blood oxygen level–dependent signals in the ventromedial prefrontal cortex and the ventral striatum predicted a higher starting point and steeper increase of the Alcohol Use Disorders Identification Test consumption score over time, respectively.

**Conclusions:** We found that MB behavioral control was associated with the binge drinking trajectory, while the MF reward prediction error signal was closely linked to the consumption

score development. These findings support the idea that the unbalanced MB and MF control might be an important individual vulnerability in predisposing to risky drinking behavior.

## 2.2 Introduction

According to a recent cross-national study, the mean lifetime prevalence of alcohol use among the world's population is 80%. The average lifetime prevalence of alcohol use disorder (AUD) is 10.7% of that population (Glantz et al., 2020), which indicates that AUD develops in only a portion of the population. Current theories about the predisposing factors of AUD point to trait impulsivity, anxiety, genetic factors, and novelty seeking along with their neural correlates (reviewed in Belin et al., 2016; Egervari et al., 2018; Jupp & Dalley, 2014). It is widely accepted that compulsive drug-seeking behavior involves a transition from choices based on action-outcome (goal-directed) to those based on stimulus-response (habitual) control (Belin et al., 2013; Everitt & Robbins, 2016; Ostlund & Balleine, 2008). The imbalance of goal-directed and habitual control frequently results in compulsive drinking behavior (Jennison, 2004), as tested in a cross-sectional design. As of yet, whether this imbalance predisposes to risky alcohol use in a longitudinal design remains untested.

Previous studies investigated how unbalanced goal-directed and habitual control was associated with compulsive drinking using the two-step sequential decision-making task (Daw et al., 2011). Developed from the reinforcement learning framework, the two-step task assesses habitual and goal-directed behavior via model-free (MF) and model-based (MB) control, respectively. To elaborate, MF control computes and updates the action value based on the reward prediction error (RPE) signal, which has been linked to the dopaminergic neurons in the midbrain (Schultz et al., 1997) and the blood oxygen level–dependent signal in the ventral striatum (VS) (Gläscher et al., 2010; O'Doherty et al., 2015; Schönberg et al., 2007). In contrast, MB control examines all possible pairs of actions and outcomes based on decision trees (Daw et al., 2005), and it is sensitive to the structure of the task (Daw et al., 2011). Accordingly, MB prediction error reflects the surprise on entering a new state given the expectation based on the task model (Gläscher et al., 2010; O'Doherty et al., 2015). To compare the two systems, MF control bases decisions on previously selected actions and is therefore inflexible, whereas MB control has more flexibility with respect to in-depth planning, but is more computationally expensive.

As evidenced by poor performance in a reversal learning task, patients with AUD were found to have an impaired MB control system. This was illustrated by behavioral deficits when challenged to integrate alternative choice options in flexible decision making (Reiter et al., 2016a). When associating AUD with the imbalance of MB and MF control, recently detoxified patients with AUD (3 weeks on average) were shown to use less MB strategy compared with healthy controls in a preliminary sample (Sebold et al., 2014). Sebold et al. (2017) further explored this topic with the full sample and attempted to predict treatment outcomes in recently detoxified patients with AUD with performance on the two-step task. Although MB behavioral control did not predict rates of relapse in the full sample, patients who relapsed showed reduced neural activation in the medial frontal cortex for MB control compared with healthy control subjects and patients who abstained. Conversely, Voon et al. (2015) examined a detoxified AUD group with varying periods of abstinence (2 weeks to 1 year) and found no differences in strategies between the AUD group and the healthy control group. Nevertheless, a link likely exists between AUD and unbalanced MB and MF control.

Similar associations were also detected in nonclinical populations. Reduced MB control has been associated with binge drinking behavior (Doñamayor et al., 2018) and the number of AUD symptoms in a large general population sample (Gillan et al., 2016). A small study (*N* = 20) did not find reduced MB control in participants with a positive family history of alcohol dependence (Reiter et al., 2016b). Even though the previously mentioned studies demonstrated an association between alcohol consumption, binge drinking, and number of AUD symptoms with unbalanced MB and MF control cross-sectionally, it is not yet clear whether an imbalance between MB and MF control predisposes to risky alcohol use and AUD or evolves from repetitive alcohol consumption. We sought to clarify whether impairments in MB reasoning are a predisposing factor for risky alcohol use using a longitudinal design and a larger sample size. We were specifically interested in early risky alcohol use and binge drinking because they typically evolve as intermediate states during the transition from occasional social drinking into compulsive alcohol use. In our study, MB and MF control were assessed by the two-step task in a community sample of 18-year-old men. Their alcohol drinking behavior was recorded over the course of 3 years, from ages 18 to 21 years, considering that alcohol consumption in this sample is legally allowed since age 16, i.e., when risky drinking behavior typically escalates (Behrendt et al., 2009; Chartier et al., 2010; Chen et al., 2004;

Muthen & Muthen, 2000). Risky drinking during this period also leads to an increased chance of developing AUD during the later stages of life (Jennison, 2004). If MB and MF control could predict risky drinking trajectory during this period, it could then be considered one of the more crucial factors that predispose to pathological drinking.

Previously, we reported no association between MB and MF control and alcohol drinking behavior at the age of 18 in this sample (Nebe et al., 2018). The current study investigated whether the two-step task performance at the age of 18 would predict the alcohol drinking developmental trajectory over the 3-year follow-up. We included both behavioral and neural predictors from the two-step task. For the neural predictors, we used both MB and MF RPE signals in the VS and ventromedial prefrontal cortex (vmPFC), as both regions have been shown to compute a mixture of the two RPE signals (Daw et al., 2011; Nebe et al., 2018). Regarding the drinking behavior, we primarily constructed two drinking trajectories with latent growth curve models: a binge drinking score assessed by the quantity of alcohol intake per drinking occasion and a consumption score assessed by the sum of the first three items of the Alcohol Use Disorders Identification Test (AUDIT). We hypothesized first that behavioral and neural correlates of MB control would be negatively associated with alcohol drinking trajectories over 3 years. Although previous studies have failed to find a clear association between MF control and AUD or risky alcohol use, a shift from MB to MF control could still be a predisposing factor—i.e., it could promote development of risky alcohol use and ultimately AUD and may not necessarily be maintained or identifiable by the time AUD has developed. Therefore, we further tested the hypothesis that behavioral and neural correlates of MF behavioral and neural control in the two-step task at baseline were associated with a steeper increase in alcohol drinking trajectory.

## 2.3 Materials and Methods

### 2.3.1 Participants & Procedure

This study was part of a longitudinal prospective study to identify learning and decision-making mechanisms underlying dysfunctional alcohol consumption during early young adulthood in a community sample (ClinicalTrials.gov identifier: NCT01744834). Only men were recruited owing to the higher prevalence of risky drinking behavior in men compared to women. The recruitment procedure and inclusion/exclusion criteria are described in

Appendix A.1. At baseline, 201 participants completed the Munich-Composite International Diagnostic Interview (Jacobi et al., 2013; Wittchen & Pfister, 1997) according to the German version of DSM-IV (Saß et al., 2003). Additionally, the participants performed the two-step task in the magnetic resonance imaging (MRI) scanner and partook in a cognitive ability assessment that examined working memory, processing speed, and crystallized intelligence (details in Appendix A.3).

Thereafter, all participants who completed the baseline assessment were invited to 6 follow-up evaluations over the course of the next 3 years. Regarding the key drinking behavior assessments, participants were asked to complete the AUDIT questionnaire online or send the completed questionnaire via post every 6 months starting from the first follow-up. However, the AUDIT questionnaire was not available for the baseline assessment at age 18. The Munich-Composite International Diagnostic Interview was conducted in person at age 18, and via telephone at ages 19, 20, and 21.

### 2.3.2 Alcohol Drinking Assessment

We constructed the drinking trajectories with two variables of interest. The average alcohol intake per drinking occasion (grams of alcohol/occasion; binge drinking score) during the past year from the Munich-Composite International Diagnostic Interview assesses the amount of alcohol consumed on a typical drinking occasion. This variable was used as a proxy for binge drinking behavior or heavy drinking episodes (Dawson, 2011; Gmel et al., 2011). The AUDIT consumption score was used as second variable to construct drinking trajectories. The AUDIT consumption score assesses the frequency of drinking, the alcohol consumption in a typical drinking occasion, and the frequency of binge drinking. Further information on the rationale of choosing these two variables and descriptive statistics are given in Appendix A.2, Table 2, and Figure S1.

In addition, we regressed the two variables against time points (modeled as categorical variables) to identify how the drinking behavior developed over the 3 years on the group level. To inspect the individual developmental trajectories, the individual intercepts and slopes (latent variables from the latent growth curve modeling [LGCM] model, which is described below) were extracted and plotted as histograms (Figure 5; C-D). The correlation between the two drinking variables was also calculated whenever they were assessed at the same time

point (at ages 19, 20, and 21). Moreover, we also tested the correlation between the two individual intercepts and slopes of the binge drinking and consumption score trajectories. These correlation tests would then indicate whether the two variables assessed different aspects of drinking behavior and followed different developmental trajectories. Descriptive statistics of additional drinking variables are displayed in Table S3.

### 2.3.3 Two-Step Paradigm

The two–step sequential decision-making task was performed in the MRI scanner (Figure 3).

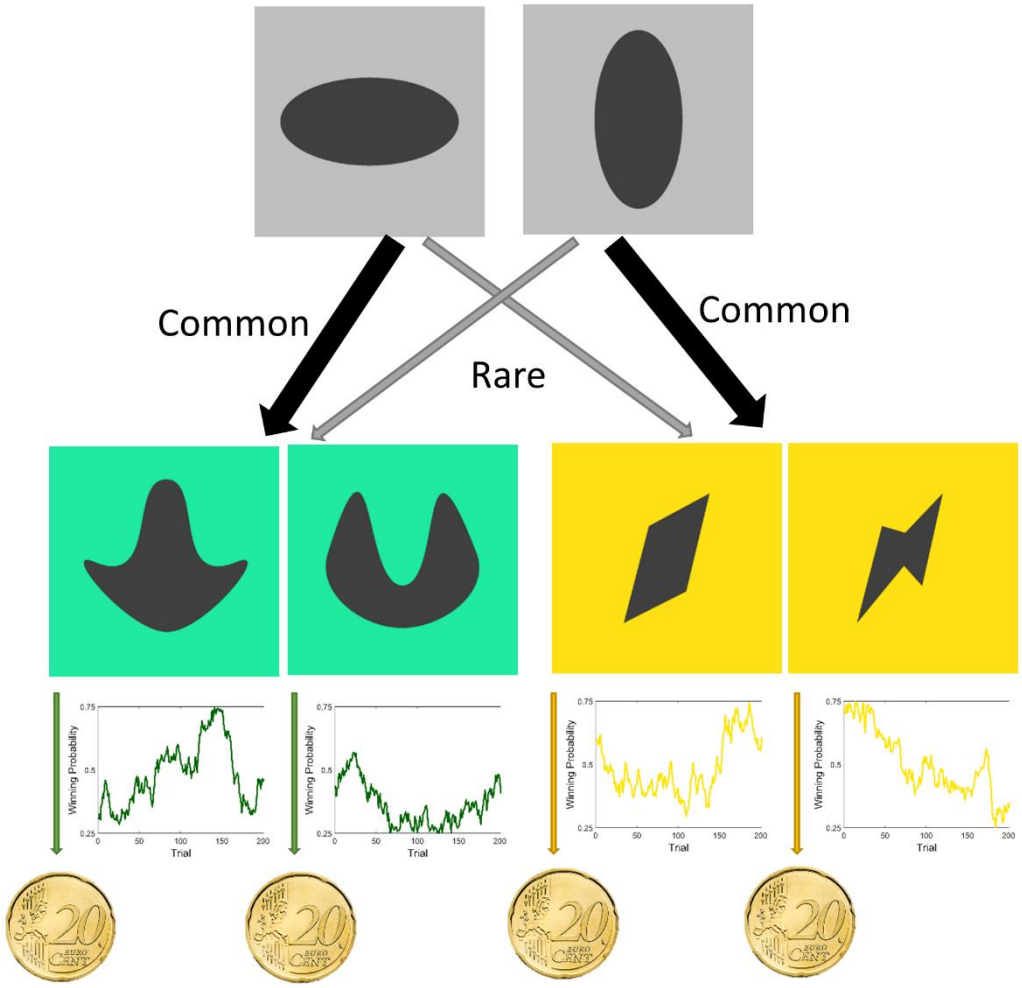


**Figure 3: The two-step paradigm.** The two-step sequential decision-making task (Daw et al., 2011) was performed in the magnetic resonance imaging scanner. The functional magnetic resonance imaging data acquisition and preprocessing procedures were described in detail in A.5 in the Appendix. The task consisted of 201 trials in total. In the first stage, the same pair of gray boxes with oval shapes inside were shown. Participants were asked to select one of these boxes within 2 seconds. The choice between the two first-stage stimuli would then lead to one of the second-stage pairs: the

common transition (with a probability of 70%) or the rare transition (with a probability of 30%). The transition probability from the first to the second stage was fixed throughout the task, and participants were informed about this. In the second stage, one of the two pairs of stimuli were presented (either yellow or green) based on the first-stage choice and the transition probability. The participants were again asked to select one of the second-stage colored stimuli within 2 seconds. The selected second-stage stimulus led to the monetary reward of 20¢ (€0.20) with a reward probability ranging from 25% to 75%, which was slowly changing across the experiment according to Gaussian random walks. In exchange for their time and cooperation, participants were paid according to the total monetary rewards acquired in one third of the trials that were randomly drawn from all trials

## 2.3.4 Two-Step Data Analysis

### *2.3.4.1 Two-Step Behavioral Predictors*

As suggested by Daw et al. (2011), who originally described the two-step task, the pure MF agent tends to ignore the structure of the task by repeating the first-stage choice after being rewarded on making their second-stage choice. Conversely, the pure MB agent considers the transition structures. The MF agent is thus sensitive to the effects of receiving a reward, as he chooses to stay after reward trials and switch after omission trials. The MB agent makes decisions based on the reward by transition interaction effect, as he tends to stay after rewarded common trials but switch after rewarded uncommon trials (and vice versa for the omission trials). It was suggested in our previous article that the participants adopted a combination of MB and MF strategy (Nebe et al., 2018). Therefore, we calculated two scores ($MB_{score}$, $MF_{score}$) for each individual according to his first-stage stay probability (P) across all trials. The purpose of these scores was to measure the degree that participants behaved like the pure MB and the pure MF agents. The two scores were then used as behavioral predictors for the alcohol drinking developmental trajectory. Specifically, they were calculated as the follows:

$MF_{score}$ = P (stay|rewarded common) + P (stay|rewarded rare) - P (stay|unrewarded common) - P (stay|unrewarded rare);

$MB_{score}$ = P (stay|rewarded common) - P (stay|rewarded rare) - P (stay|unrewarded common) + P (stay|unrewarded rare).

*2.3.4.2 Two-Step Neural Predictors*

A total of 146 participants were included in the final fMRI analysis after quality control [same as in Nebe et al. (2018)]. The fMRI first-level model is the same as our baseline report; one onset regressor was specified for the second-stage onset, with MB RPE and MF RPE modeled as two parametric modulators (see details in Appendix A.6). To assess the neural correlates of MB and MF RPE signal, we performed one-sample *t* tests on both MB and MF RPE parametric regressors on the second level. Consistent with previous studies (Daw et al., 2011; Nebe et al., 2018), two regions of interest were specified: the bilateral vmPFC and bilateral VS (based on meta-analyses; see A.6 in the Appendix). Both the VS and vmPFC have been suggested to compute a mixture of MB and MF RPE signals, and these cannot be disentangled in the two-step task (Daw et al., 2011). It was for this reason that the mean parameter estimates within the two regions of interest were extracted separately for both MB and MF RPE parametric regressors. The four neural predictors were then applied to predict the alcohol drinking developmental trajectory.

### 2.3.5 LGCM Analysis

LGCM offers an elegant framework to model both intra- and interindividual change over time (Duncan & Duncan, 2009). Traditional approaches, such as analysis of variance, treat individual differences as variances. Unlike analysis of variance, though, LGCM additionally models the intraindividual change. As a multilevel model, intraindividual change in drinking behavior with respect to time was modeled on the first level. Thus, one intercept and one slope can characterize an individual's drinking behavior when a linear developmental trajectory is assumed. Different individual drinking developmental trajectories can be identified accordingly. Based on our hypothesis, individual drinking trajectories were modeled for the aforementioned variables: gram/occasion and AUDIT consumption score. Additionally, a model comparison was performed between quadratic and linear models to decide if adding a quadratic term to the model would improve the model fit. It was found that the model fit of the quadratic trajectory models was worse than the linear trajectory models (see details in Appendix A.7). Therefore, we tested only the predictors from the two-step task in the linear trajectory models.

At the second level, predictors of interest could be included in the model. The model would then decide whether these predictors are associated with the interindividual differences in their developmental trajectories, as indicated by the individual intercepts and slopes from the first level (when the linear trajectory model fits better than the quadratic model). Our model included six predictors of interest: MB and MF behavioral scores and MB and MF neural signals in the VS and the vmPFC, respectively. Freely estimated covariances were allowed between MB and MF behavioral scores, between MB neural signals, and between MF neural signals. The two models of interest are displayed in Figure 4. In addition, executive functions (Schad et al., 2014) and impulsivity (Deserno et al., 2015; Reiter et al., 2016b) are thought to be associated with two-step task performance. We also found associations between the MB behavioral control, working memory capacity, and processing speed in our sample (A.3 and Table S2 in the Appendix), but no association was found between the two-step predictors and impulsivity measured by the Barratt Impulsiveness Scale sum score (Table S6 in the supplementary information in Nebe et al., 2018). Nevertheless, to control for the potential effect of these factors on the models, these variables were included separately in the two models to check whether they had an effect. Additionally, because we previously reported that low MB control is associated only with increased risk for relapse in patients with AUD with high alcohol expectancies (Sebold et al., 2017), we explored whether such an interaction between MB control and alcohol expectancies also existed in our sample. The detailed analyses and results are shown in A.11 in the Appendix.

### 2.3.6 LGCM Model Structure and Path Estimates

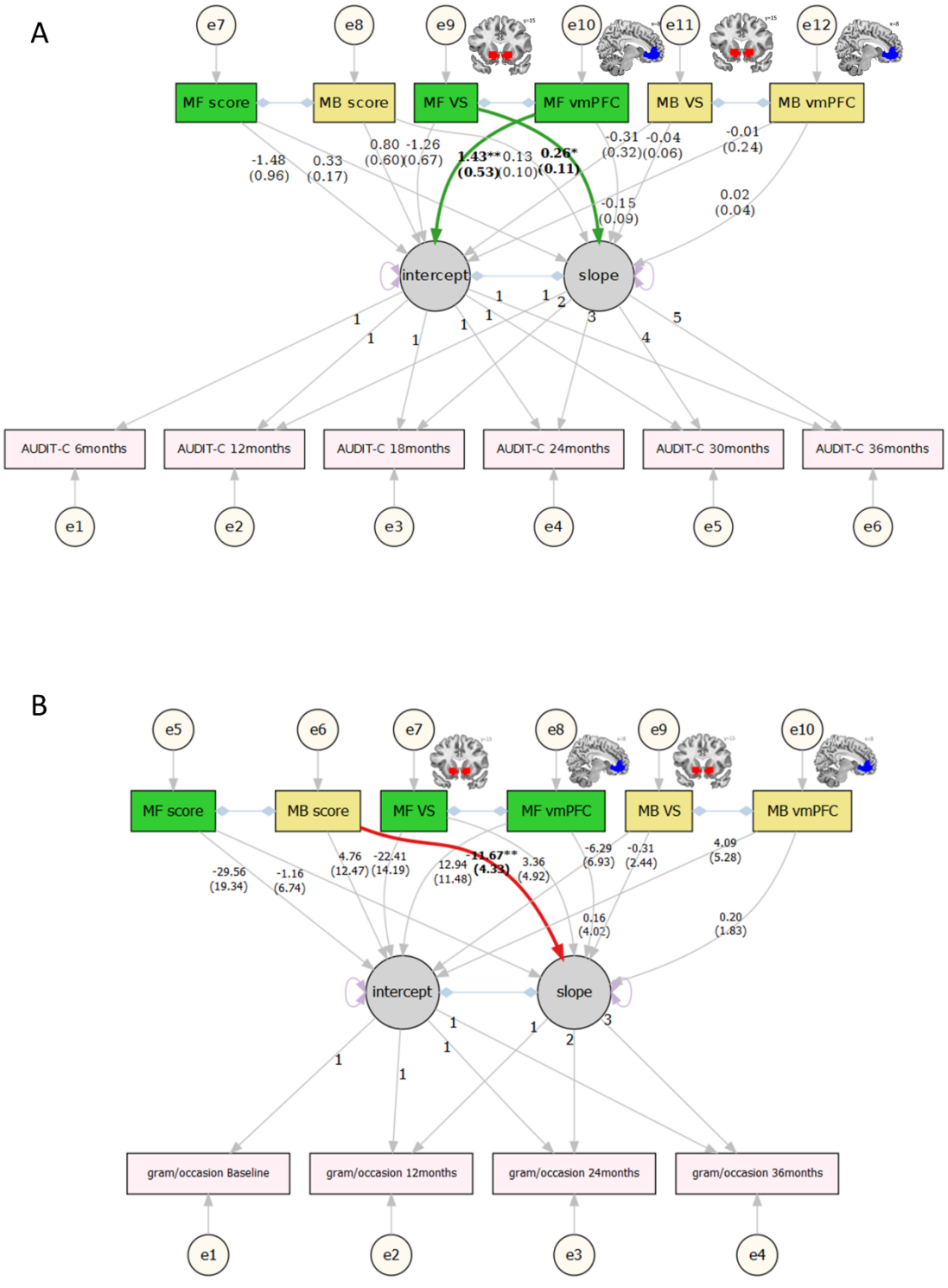


**Figure 4: Latent growth curve modeling structure.** Alcohol Use Disorders Identification Test consumption score (AUDIT-C) model (A) and gram/occasion model (B). The intercept and slope were modeled as the latent variables. All the other variables were observed from the data. The loadings

from the intercept and slope to the drinking variables were fixed with values shown in the figure, indicating the linear trajectory. All the other paths including regressions, variances, and covariances were freely estimated from the model. Latent growth curve modeling was modeled within the structural equation model framework using the lavaan package in R (Rosseel, 2012). The lavaan package allows for the handling of missing data with full information maximum likelihood, which estimates a likelihood function for each individual based on the available information. This method is suggested to be unbiased, as the missing data are assumed to be random (Arbuckle et al., 1996). For path estimation, as indicated by the green paths, the model-free (MF) ventromedial prefrontal cortex (vmPFC) and the MF ventral striatum (VS) signals were positively associated with the intercept and slope, respectively, in the AUDIT-C model; the model-based (MB) behavioral score was negatively associated with the slope of the gram/occasion (binge drinking) model. The standardized estimates are displayed in Table 1.

## 2.4 Results

### 2.4.1 Drinking Trajectories

According to the two linear regressions of the two drinking scores against time points, the AUDIT consumption score did not change with time on the group level ($\beta = -0.06$; $p = .26$), but the binge drinking score (gram/occasion) significantly decreased over time ($\beta = -8.54$; $p = 3.81 \times 10^{-7}$). However, as can be seen in the trajectory plots and the histograms of individual intercepts and slopes (Figure 5), individuals exhibited different developmental trajectories within the 3-year time course even without overall significant changes. The two drinking scores correlated with each other early on but tended to develop independently over time (correlations shown in Figure 5).

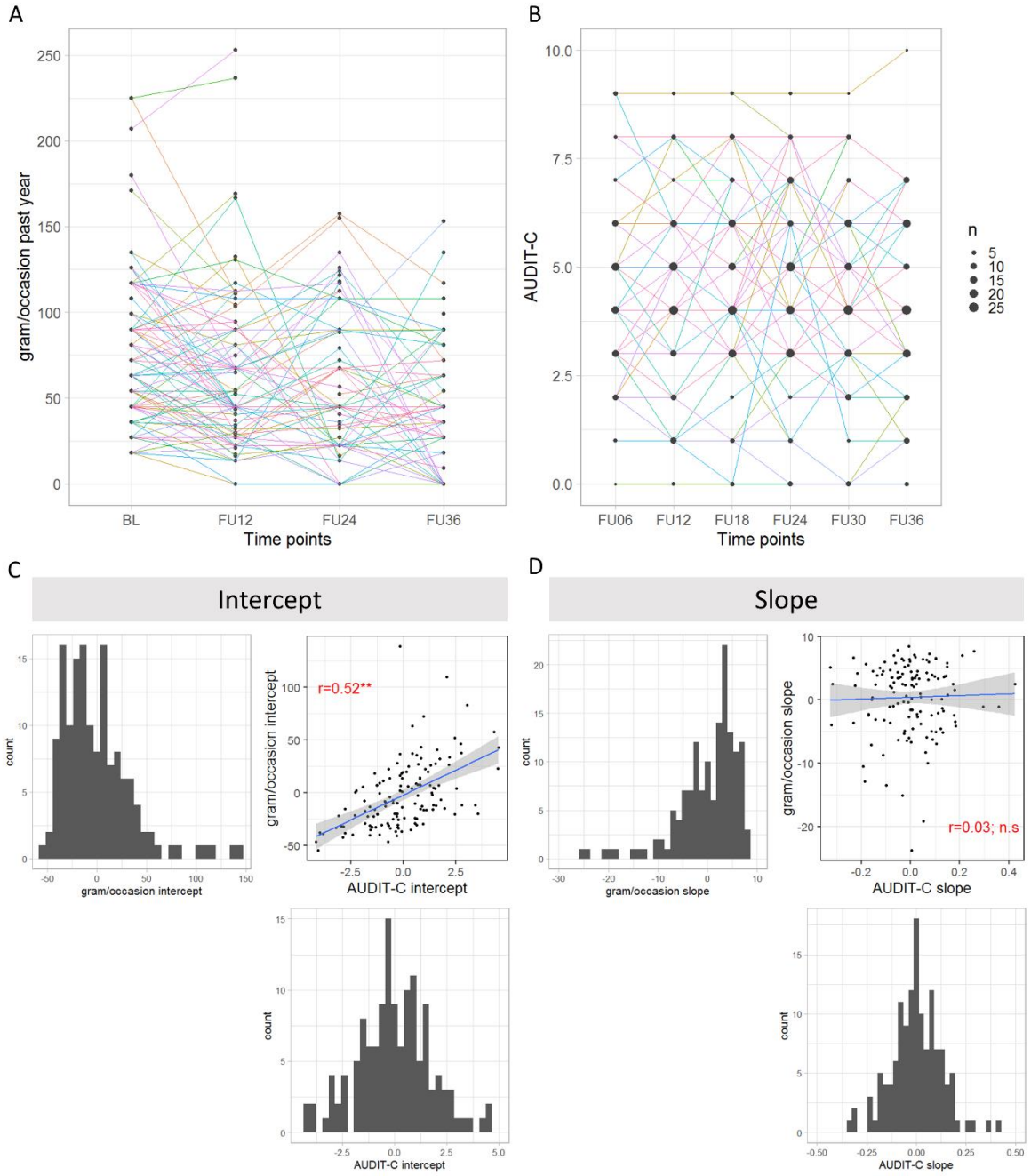


**Figure 5. Individual drinking trajectories.** (A, B) Individual trajectories (indicated by different colors) of gram/occasion and Alcohol Use Disorders Identification Test consumption score (AUDIT-C) variables across different time points. Individual differences in the developmental trajectories within the 3 years can be seen. The measure of gram/occasion during the last year yielded four time points; the AUDIT-C was assessed every 6 months after baseline (BL) and yielded 6 time points (FU06, FU12, FU18, FU24, FU30, FU36). A total of 146 participants with valid data were included in the gram/occasion trajectory. We further excluded 13 participants who lacked valid AUDIT assessments over the 3 years. The correlation between the gram/occasion variable and AUDIT-C was moderate at the age of 19 ($r_{77}$ = .49,

$p = 5.07 \times 10^{-6}$), strong at the age of 20 ($r_{74} = .61$, $p = 4.58 \times 10^{-9}$), but low at the age of 21 ($r_{97} = .29$, $p = 3.46 \times 10^{-2}$). (C, D) The individual intercepts from the two drinking models showed a significant correlation ($r_{131} = .52$, $p = 1.35 \times 10^{-10}$), but the 2 slopes were not correlated ($r_{131} = .03$, $p = .73$). n.s., not significant.

### 2.4.2 LGCM Model Results

The AUDIT consumption score model (Figure 4A) demonstrated a good model fit ($\chi 2_{48} = 81.12$, $p = .002$, comparative fit index = 0.956, root mean square error of approximation = 0.072, standardized root mean square residual = 0.078). The path parameter estimates are shown in Table 1. Among the predictors, we found that the MF VS signal was positively associated with a change in AUDIT consumption score over time (slope), while MF vmPFC activation was positively associated with AUDIT consumption score in the 6-month follow-up (intercept). The association between MF behavioral score and slope was also positive, but this effect was only marginal ($p = .055$).

The binge drinking model displayed in Figure 4B showed a good model fit as well ($\chi 2_{27} = 50.26$, $p = .004$, comparative fit index = 0.935, root mean square error of approximation = 0.077, standardized root mean square residual = 0.084). As displayed in Table 1, we found that the two-step MB behavioral score was negatively associated with the developmental trajectory (slope) of the gram/occasion variable over the past year. The four neural predictors and the MF behavioral score did not show significant associations with either the intercept or the slope of the gram/occasion during the last year. Additional exploratory analyses with alcohol expectancies showed that only individuals with high expectations of the positive reinforcing effect of alcohol showed the negative association (see details in A.11 in the Appendix).

*Table 1: LGCM results*

| | | Path | Estimate (unstandardized) | SE | Estimate (standardized) | *z* | *p* value | Effect size[a] ($r^2$) |
|---|---|---|---|---|---|---|---|---|
| **AUDIT consumption score** | | | | | | | | |
| intercept | MF | Behavioral score | -1.476 | 0.960 | -0.156 | -1.537 | .124 | 2.4 % |
| | | VS signal | -1.257 | 0.669 | -0.239 | -1.880 | .060 | 5.7 % |
| | | vmPFC signal | 1.428 | 0.534 | 0.341 | 2.675 | **.007**[b] | 15.3 % |
| | MB | Behavioral score | 0.801 | 0.600 | 0.139 | 1.337 | .181 | 3.6 % |
| | | VS signal | -0.307 | 0.322 | -0.131 | -0.952 | .341 | 1.7 % |
| | | vmPFC signal | -0.011 | 0.243 | -0.006 | -0.044 | .965 | 0.0 % |
| slope | MF | Behavioral score | 0.327 | 0.171 | 0.302 | 1.918 | .055 | 12.4 % |
| | | VS signal | 0.259 | 0.113 | 0.429 | 2.286 | **.022**[c] | 22.9 % |
| | | vmPFC signal | -0.151 | 0.093 | -0.314 | -1.628 | .104 | 9.9 % |
| | MB | Behavioral score | 0.131 | 0.104 | 0.198 | 1.268 | .205 | 6.2 % |
| | | VS signal | -0.035 | 0.057 | -0.130 | -0.618 | .537 | 1.7 % |
| | | vmPFC signal | 0.018 | 0.042 | 0.090 | 0.419 | .675 | 2.0 % |
| **Binge drinking score (Gram alcohol/Drinking Occasion) Past Year** | | | | | | | | |
| intercept | MF | Behavioral score | -29.560 | 19.335 | -0.151 | -1.529 | .126 | 2.3 % |
| | | VS signal | -22.413 | 14.188 | -0.202 | -1.580 | .114 | 4.1 % |
| | | vmPFC signal | 12.944 | 11.481 | 0.144 | 1.127 | .260 | 3.8 % |
| | MB | Behavioral score | 4.755 | 12.465 | 0.039 | 0.381 | .703 | 0.8 % |
| | | VS signal | -6.291 | 6.927 | -0.129 | -0.908 | .364 | 1.7 % |
| | | vmPFC signal | 4.094 | 5.278 | 0.113 | 0.776 | .438 | 2.7 % |
| slope | MF | Behavioral score | -1.159 | 6.735 | -0.030 | -0.172 | .863 | 0.1 % |
| | | VS signal | 3.359 | 4.918 | 0.152 | 0.683 | .495 | 4.1 % |
| | | vmPFC signal | 0.157 | 4.017 | 0.009 | 0.039 | .969 | 0.3 % |
| | MB | Behavioral score | -11.662 | 4.329 | -0.483 | -2.694 | **.007**[b] | 23.3 % |
| | | VS signal | -0.309 | 2.435 | -0.032 | -0.127 | .899 | 0.1 % |
| | | vmPFC signal | 0.196 | 1.828 | 0.027 | 0.107 | .915 | 0.6 % |

** P value < .05 ** P value < .01*

AUDIT consumption score model fit: $\chi 2_{48} = 81.12$, $p = .002$, CFI = 0.956, RMSEA = 0.072, SRMR = 0.078; Binge drinking score past year model fit: $\chi 2_{27} = 50.26$, $p = .004$, CFI = 0.935, RMSEA = 0.077, SRMR = 0.084.

AUDIT, Alcohol Use Disorders Identification Test; CFI, comparative fit index; MB, model-based; MF, model-free; LGCM, latent growth curve modeling; RMSEA, root mean square error of approximation; SRMR, standardized root mean square residual; vmPFC, ventromedial prefrontal cortex; VS, ventral striatum.

[a] Effect size is displayed as the percent of explained variance ($r^2$). Correlation coefficients ($r$) were converted from the standardized coefficient according to Peterson and Brown (2005) by using the equation $r = \beta + .05\lambda$, where $\lambda$ equals 1 when $\beta > 0$ and equals 0 when $\beta < 0$.

[b] $p < .01$.

[c] $p < .05$.

Additionally, the individual latent intercepts and slopes were extracted from the two models and plotted against the significant predictors for the purpose of illustration (Figure 6). The control variables—executive functions and impulsivity score—neither changed the model estimates nor showed significant associations with the intercepts or slopes. The detailed results are shown in Table S5 and Table S6.

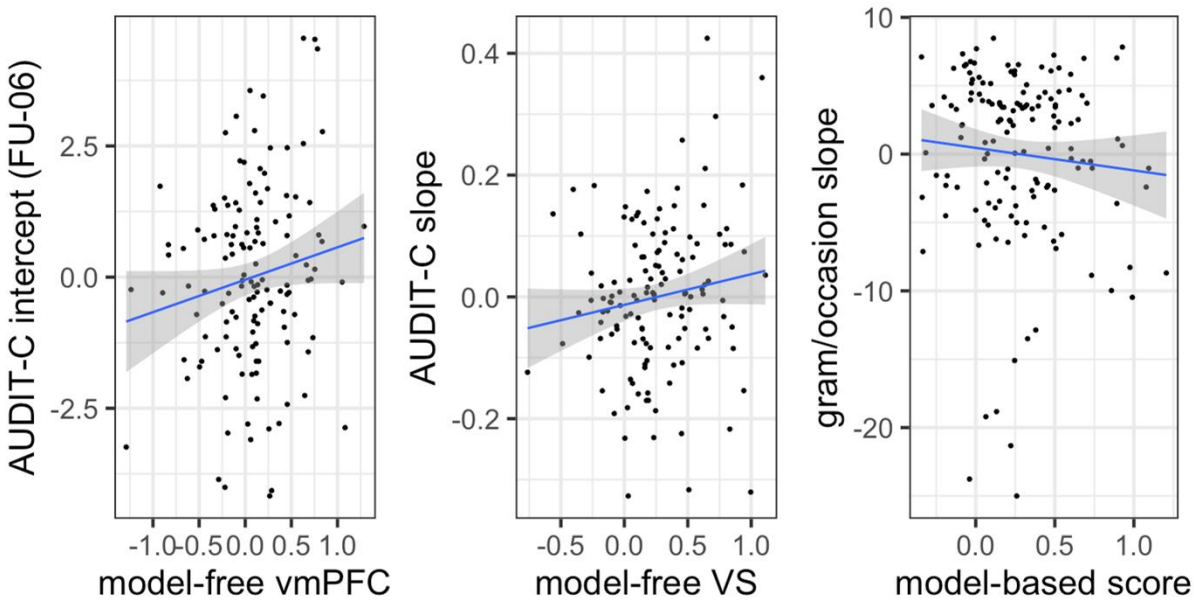


**Figure 6: Illustration of the significant paths from latent growth curve modeling**. The model-free neural RPE signal in the ventral striatum (VS) predicted higher Alcohol Use Disorders Identification Test consumption score (AUDIT-C) intercept (6 months following the baseline, FU-06). The model-free neural reward prediction error in the ventromedial prefrontal cortex (vmPFC) predicted an increase/less decrease of AUDIT-C over the 2.5 years. The model-based behavioral score was negatively associated with the slope of the gram/occasion variable.

## 2.5 Discussion

With a large community sample, we found that an unbalanced MB and MF control assessed by the two-stage sequential decision-making task at the age of 18 predicted the developmental trajectories of the binge drinking and the consumption scores during young adulthood. Specifically, MB behavioral control was associated with less increase or more decrease in the developmental trajectory of binge drinking behavior. Concerning the consumption score assessed by the AUDIT questionnaire, the neural MF RPE signal in the vmPFC and the VS predicted a higher starting point and steeper increase/flatter decrease over

time, respectively. All the identified associations had medium effect sizes (explaining 15% - 23% of variance). We thus conclude that a bias away from MB and towards MF control may represent a critical mechanism predisposing toward risky alcohol drinking during young adulthood.

Interestingly, we found that MB and MF control predict different aspects of drinking behavior. The binge drinking trajectory (i.e., slope of gram/occasion variable) was negatively associated with the MB behavioral score. Binge drinking has recently been related to deficits in executive functions, such as poor inhibitory control during adolescence and young adulthood (Carbia et al., 2018; Lees et al., 2019). Moreover, young binge drinkers are comparable to patients with severe AUD in their executive control abilities (Lannoy et al., 2019); binge drinking has also been suggested to be a consequence of the effect of alcohol on the brain networks underlying inhibitory control in young adults (Gan et al., 2014). Consistent with a previous study (Schad et al., 2014), the MB behavioral score in our sample was also associated with several facets of executive function, including processing speed, working memory capacity, and verbal intelligence. However, these executive functions per se neither predicted the drinking trajectory nor affected the model estimates. Taken together, the MB score may be closely linked to executive function but explains additional variance of binge drinking behavior. It is worth mentioning that in our sample binge drinking decreased between the ages of 18 to 21. High MB control may work as a protective mechanism by further decreasing binge drinking over time.

Notably, the MB neural signal was not associated with binge drinking behavior. On one hand, this may be due to the noise in the neural signals, which might not reliably capture the trial-by-trial MB control. On the other hand, MB behavior was not necessarily guided by the MB RPE defined in the current computational model, which was also pointed out by Daw et al. (2011). The MB control in the current task tracks transition probabilities and immediate rewards. Another way of defining the MB control posits that the state prediction error signal can be tracked by examining future planning and calculating cumulative future rewards (Gläscher et al., 2010; O'Doherty et al., 2015). However, whether this type of MB prediction error signal was computed or associated with the MB behavioral control cannot be tested with the current task design. Owing to this discrepancy, the MB behavioral predictor may

capture different aspects of the MB control and predict future binge drinking behavior better than the neural signals.

The consumption score trajectory, assessed by the first three items of the AUDIT questionnaire, was predicted by the MF RPE signals in the VS and the vmPFC. Additionally, a weaker positive association between the MF behavioral score and the development (i.e., slope) of risky alcohol use was identified at a trend level. So far, only a limited number of factors associated with alcohol consumption have been identified that are not specific to binge drinking. Two cross-sectional electroencephalography studies have found associations between higher alcohol consumption with attenuated feedback-related negativity amplitudes (Cao et al., 2021; Soder et al., 2019) and a feedback-locked P3 component (Cao et al., 2021). These two event-related components indicated the RPE signals after receiving rewards. Intriguingly, neither the feedback-related negativity nor the P3 component was found to be related to binge drinking behavior when tested with the same balloon analog risk task (BART) as in Soder et al. (2019). Therefore, we propose that the consumption score, but not specifically binge drinking, may be associated with aberrant RPE processing in the brain. In line with this, higher gray matter volume in the caudate nucleus at age 14 was found to predict a steeper increase in AUDIT score over a 5-year period (Kühn et al., 2016). Although the intercept of the binge drinking and the AUDIT consumption score were positively correlated, we did not observe an association between the MF vmPFC signal and the binge drinking intercept. This suggests that the frequency of drinking as well as of binge drinking assessed with the AUDIT in addition to mere amount of drinking per occasion may also play an important role.

Essentially, our results were intrinsically consistent, though indicated by different parameters. Lower MB or higher MF control indicated riskier drinking trajectories. As discussed above, MB and MF control seem to predict different facets of drinking behavior. The associations between MB control and drinking were significant only for the behavioral indicator, whereas the association between MF control were significant only for the respective neural signatures. One explanation is that some predictors might not have been identified because effect sizes are smaller (e.g., MF behavioral score) or are due to lower reliability compared to the ones identified. The MB neural signal, for example, may be noisy, which means that an even larger sample would be required to discern any effects. Furthermore, MB control could also be

promoted by more detailed task instructions (da Silva & Hare, 2020). Therefore, when participants misconceive the task, there might also be a mismatch between the strategies that participants used and the strategy captured by the model. Taken together, larger sample sizes, an improved version of the paradigm (Kool et al., 2016), and improved parameter estimates (Shahar et al., 2019a) may potentially resolve such discrepancies.

## 2.6 Limitations

Although MB and MF control were found to predict risky drinking during the 3-year follow-up, we do not have any information about whether the participants with risky drinking trajectories would develop AUD in a later phase of life. This would require a longer follow-up period, as direct evidence is needed. Additionally, we assumed that the missing data are at random, but we could not test whether other factors contributed to participants dropping out of the study. Given that the missing rates are at 30%-40% at almost every time point, computational methods had to be applied to preserve the data. Nevertheless, we did not reach the current conclusions without the assumptions about the missing data. Also, the AUDIT was first assessed 6 months after the baseline. We thus could not infer the association between the two-step predictors and general risky alcohol consumption at baseline, but rather only 6 months later. Lastly, this study included only male participants, and therefore the results cannot be generalized to non-male populations.

## 2.7 Conclusions

By assessing two modes of instrumental learning (i.e., MB and MF learning) and recording the drinking behavior of a large cohort of young men over a period of 3 years, we were able to identify predictors of risky alcohol use. Our data reveal that a higher MB behavioral score predicts a decrease in binge drinking, while a higher MF RPE neural signal predicts a higher AUDIT consumption score that further increases over time. Our findings may also suggest that the AUDIT consumption score and binge drinking trajectories may develop differently during young adulthood and involve different mechanisms. Dysbalanced control might ultimately also predispose for the later development of AUD, but the duration of the follow-up and the limited sample size do not allow drawing conclusions yet. We propose that future studies could further examine these links by carefully assessing different aspects of alcohol consumption in larger cohorts and over longer periods of time. To better comprehend the

link between the unbalanced MB and MF control and (pathological) alcohol use, another direction for future studies is to investigate the consequences of drinking, i.e., whether alcohol consumption further changes the MB and MF control. Lastly, the current study also opened a new door for future studies to develop interventions to target these proposed mechanisms in preventing risky alcohol use.

## 2.8 Acknowledgements and Disclosures

This study was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft) (Grant Nos. 186318919 [FOR 1617, to AH, MAR, and MNS], 178833530 [SFB 940, to MNS], and 402170461 [TRR 265, to AH, MAR, and MNS]), and University of Zurich Grants Office (Grant No. FK-19-020 [to SN]).

SKP received honoraria/fees during the past 12 months: author fees from a publisher of medical books (Mabuse Verlag), and honoraria for one speech from a group of companies (AbbVie Deutschland, Almirall Hermal, Belano medical, Celgene, Janssen-Cilag, LEO Pharma, Lilly Deutschland, Novartis Pharma, Pfizer Pharma, and UCB Pharma). QJMH has received consultancy fees from Aya Technologies. All other authors report no biomedical financial interests or potential conflicts of interest.

# Chapter 3: PIT and Risky Drinking at Age 18 (Study 2)

This chapter has been published as the following:

**Chen, H**., Nebe, S., Mojtahedzadeh, N., Kuitunen-Paul, S., Garbusow, M., Schad, D. J., Rapp, M. A., Huys, Q. J. M., Heinz, A. & Smolka, M. N. (2021). Susceptibility to interference between Pavlovian and instrumental control is associated with early hazardous alcohol use. *Addiction biology*, *26*(4), e12983.

## 3.1 Abstract

Pavlovian-to-instrumental transfer (PIT) tasks examine the influence of Pavlovian stimuli on ongoing instrumental behavior. Previous studies reported associations between a strong PIT effect, high-risk drinking, and alcohol use disorder. This study investigated whether susceptibility to interference between Pavlovian and instrumental control is linked to risky alcohol use in a community sample of 18-year-old male adults. Participants (*N* = 191) were instructed to “collect good shells” and “leave bad shells” during the presentation of appetitive (monetary reward), aversive (monetary loss), or neutral Pavlovian stimuli. We compared instrumental error rates (ER) and fMRI brain responses between the congruent and incongruent conditions, as well as among high-risk and low-risk drinking groups. On average, individuals showed a substantial PIT effect, that is, increased ER when Pavlovian cues and instrumental stimuli were in conflict compared with congruent trials. Neural PIT correlates were found in the ventral striatum and the dorsomedial and lateral prefrontal cortices (lPFC). Importantly, high-risk drinking was associated with a stronger behavioral PIT effect, a decreased lPFC response, and an increased neural response in the ventral striatum on the trend level. Moreover, high-risk drinkers showed weaker connectivity from the ventral striatum to the lPFC during incongruent trials. Our study links interference during PIT to drinking behavior in healthy, young adults. High-risk drinkers showed higher susceptibility to Pavlovian cues, especially when they conflicted with instrumental behavior, indicating lower interference control abilities. Increased activity in the ventral striatum (bottom-up), decreased lPFC response (top-down), and their altered interplay may contribute to poor interference control in the high-risk drinkers.

## 3.2 Introduction

To behave efficiently in one's daily life and to adapt one's actions to a dynamic environment, a response selection system is frequently engaged. Critical control components involved when making such choices include Pavlovian and instrumental control. Through Pavlovian conditioning, inborn and hard-wired responses (e.g., approach or avoidance) to biologically potent (unconditioned) stimuli can be associated with neutral stimuli. Thereafter, such conditioned responses to Pavlovian cues are independent of their outcomes. Conversely, instrumental behavior, more specifically, goal-directed instrumental behavior, is controlled by the contingencies between actions and outcomes. Pavlovian cues can influence ongoing instrumental behavior, even though the responses to the Pavlovian cues were acquired separately from the instrumental responses—this process is called Pavlovian-to-instrumental transfer (PIT). To elaborate, a food's enticing scent (Pavlovian) may encourage people to partake in eating behavior (instrumental), whereas an unpleasant scent may hinder the eating behavior. In a typical human PIT task (Cartoni et al., 2016; Holmes et al., 2010), participants need to perform learned instrumental responses (press a button for approach or avoidance) in the presence of previously and independently trained Pavlovian cues (appetitive or aversive).

Most previous human PIT studies investigated how Pavlovian cues influence instrumental approach behavior. Accordingly, appetitive Pavlovian cues were found to promote instrumental approach responses compared to the neutral cues (Allman et al., 2010; Eder & Dignath, 2016a, 2016b; Paredes-Olay et al., 2002; Quail et al., 2017; Rosas et al., 2010; Watson et al., 2014), whereas aversive Pavlovian cues were found to reduce instrumental approach behavior (Geurts et al., 2013; Huys et al., 2011). Additionally, some studies have examined PIT effects in the avoidance context by rewarding successful instrumental avoidance behavior, in which aversive Pavlovian cues were shown to promote instrumental avoidance behaviors (Garofalo & Robbins, 2017; Lewis et al., 2013; Nadler et al., 2011).

Moreover, in an orthogonal experimental design with the appetitive – aversive Pavlovian axis and the approach – avoidance instrumental axis, instrumental behavior was impaired by incongruent Pavlovian cues (instrumental approach behavior by aversive Pavlovian cues, or instrumental avoidance behavior by appetitive Pavlovian cues) but was promoted by

congruent Pavlovian cues (Huys et al., 2011; Sommer et al., 2017). Freeman et al. (2015) used a go-no-go/PIT task which resembles a classical go-no-go task. In this task, participants learned to respond to one stimulus in the go trials while withholding their responses to another stimulus in no-go trials. The authors modified the proportion of no-go trials where appetitive Pavlovian cues were presented. It was then found that when the proportion of incongruent no-go trials out of all no-go trials was higher, the provocation of the appetitive cues on instrumental approach behavior (go trials) in the subsequent trials was reduced. Additionally, in one EEG study, Cavanagh et al. (2013) used another variant of a go-no-go task to investigate how Pavlovian biases influence instrumental learning during the conflict between both systems. It was found that midfrontal theta power, sensitive to conflict and the following adaptive control, was associated with the ability to overcome Pavlovian biases when they interfered with the instrumental behavior. Taken together, these four studies imply that cognitive control is to be allocated to overcome the conflict between Pavlovian and instrumental control.

Linked to alcohol drinking behavior, previous studies from our group have found associations between the stronger motivational effect of Pavlovian cues on instrumental behavior and alcohol dependence (Garbusow et al., 2016; Garbusow et al., 2014; Schad et al., 2019), as well as high-risk drinking during young adulthood (Garbusow et al., 2019). In addition to the enhanced behavioral effect, the neural correlates of the motivational PIT effect in the nucleus accumbens (Garbusow et al., 2014; Schad et al., 2019) and the amygdala (Garbusow et al., 2019) were also associated with alcohol dependence and high-risk drinking during young adulthood, respectively. Notably, when whether the Pavlovian cue interferes with the instrumental behavior was taken into account, alcohol-dependent patients committed more errors compared with healthy controls when Pavlovian stimuli and instrumental responses were in conflict, especially when participants needed to inhibit instrumental approach responses during the presence of appetitive Pavlovian cues (Sommer et al., 2017); this behavioral impairment was also stronger for future relapsers (Sommer et al., 2020). As of yet, whether this interference effect along with its neural correlates were associated with high-risk drinking during young adulthood is not clear.

We thus investigated interference control during a PIT task in a group of healthy, young men at age 18, who were drinking occasionally but did not fulfil the criteria for Diagnostic and

Statistical Manual of Mental Disorders (DSM-IV) alcohol dependence. The rationale behind this is that social drinking behavior is influenced by numerous environmental cues during social occasions, which reflects the PIT task in the experimental settings to some extent. A reduction in the ability to allocate cognitive resources in order to control the response to cues that look tempting but violate the long-term goals may contribute to hazardous drinking development. From this perspective, we assumed that the ability to allocate interference control when the Pavlovian cues conflict with the instrumental behavior, along with its associated neural responses, could be potential (bio)markers of hazardous drinking behavior during early adulthood. More specifically, on the behavioral level, it was hypothesized that error rates (ERs) would increase in the incongruent condition, that is, when Pavlovian cues and instrumental stimuli are incongruent, as compared with the congruent condition. Importantly, individuals with higher levels of risk in drinking should show more susceptibility to this effect, that is, show lower interference control.

On the neural level, previous literature has found neural correlates of motivational effects of Pavlovian cues in the amygdala (Geurts et al., 2013; Mendelsohn et al., 2014; Prevost et al., 2012; Talmi et al., 2008), the ventral striatum (VS) (Geurts et al., 2013; Mendelsohn et al., 2014; Talmi et al., 2008) and the dorsal striatum (Bray et al., 2008; Lewis et al., 2013). Accordingly, the VS and amygdala were expected to show responses during the PIT task. Importantly, referring to the meta-analysis of tasks that require different dimensions of inhibitory or interference control (Hung et al., 2018), we also hypothesized that conflict between Pavlovian cues and required instrumental behavior would elicit responses in cognitive control areas—the lateral prefrontal cortex (lPFC) and the dorsomedial prefrontal cortex (dmPFC). Further, low-risk drinkers were hypothesized to allocate more top-down interference control as compared with high-risk drinkers. If this were to be the case, we would expect the effective connectivity between the aforementioned brain regions to be altered in the high-risk drinkers, which we would explore with dynamic causal models.

## 3.3 Materials and Methods

### 3.3.1 Participants & General Procedure

Invitation letters were first sent to 1,937 males at age 18 who were randomly sampled from the local registration offices in Dresden and Berlin, Germany. At the baseline of the

longitudinal study, only males were recruited because of the higher prevalence of risky drinking behavior. After screening 445 respondents, those with the inclusion criteria of right-handedness, no history of major mental disorders including substance dependence (except for nicotine dependence), eligibility for magnetic resonance imaging (MRI), and having had at least two drinking occasions in the past 3 months were further invited. Of those who met the inclusion criteria, 201 participants completed the behavioral and MRI assessment. After excluding participants with incomplete behavioral data because of technical issues, 191 participants were included for the final analysis.

Participants went through the experimental procedure with two appointments. During the first appointment, participants finished the Munich Composite International Diagnostic Interview (M-CIDI [Jacobi et al., 2013; Wittchen & Pfister., 1997]) according to the DSM-IV (Saß et al., 2003), along with cognitive ability assessment (details in Appendix B.2). The risk status of our subjects was defined according to their binge drinking behavior based on World health Organization standards (Stockwell et al., 2000): as recommended, an average intake of more than 60 g of ethanol on a given drinking occasion was used as a cut-off for high-risk and low-risk drinkers. According to the self-reported alcohol intake per occasion during the last year reported in the M-CIDI, 97 participants were classified as low-risk drinkers, and the other 94 as high-risk drinkers (drinking behaviors of the two groups shown in Table 2).

During the second appointment, approximately 9 days (standard deviation = 16 days) later, participants performed the PIT task consisting of four phases. The Pavlovian phase and the PIT phase were done within the MRI scanner, whereas the instrumental phase and the forced-choice phase were conducted outside the scanner. As briefly mentioned above, participants were presented with images of various shells whose quality (good or bad) was randomly assigned. During the instrumental training, participants were asked to learn the quality of each shell through trial-and-error instrumental responses. When collecting or leaving the shells, the participants received probabilistic feedback that dictated whether their action resulted in a monetary gain or loss. To collect a shell, the participants were required to press the left mouse button five or more times. Each button press resulted in a visual cue (a small dot) moving closer to the image of the shell. To leave a shell, there was no action required. A shell was only considered “collected” if the threshold of five button presses was reached or surpassed. During the Pavlovian conditioning, participants passively learned the association

between five types of compound conditional stimuli (CSs, consisting of fractal-like images and pure tones) and positive (€1, €2), negative (€-1, €-2) or neutral (€0) unconditioned stimuli (USs). Following this, participants performed the instrumental task again (90 trials) with the fractal images of the CSs tiled in the background. This phase, referred to as the PIT phase, was performed under nominal extinction to avoid further learning. Additionally, there were 72 trials with pictures of alcoholic/water beverages presented in the background in combination with the two instrumental stimuli (details about the alcohol/water PIT trials shown in Appendix B.1). In the last phase, participants were presented with two CSs within 2 s and were required to choose one. A more detailed PIT task description is provided in Figure 7 (also see Garbusow et al. [2014b]). Participants also rated their subjective experience with the five Pavlovian fractals after the experiment. The analyses for the subjective ratings and forced-choice query trials are presented in Appendix B.6.

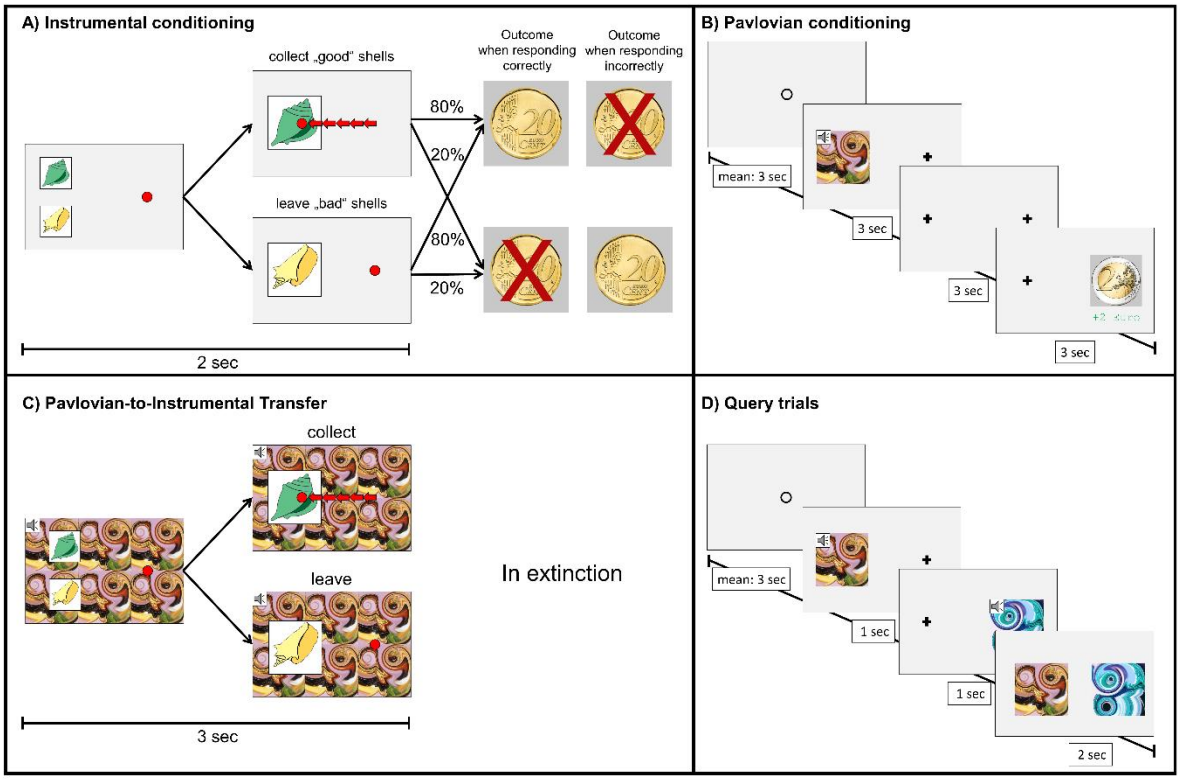


**Figure 7: Pavlovian-to-instrumental transfer (PIT) experiment procedure** (also see Garbusow et al., 2016; Garbusow et al., 2014). (A) Instrumental Phase: Participants learned to collect the good shells (press the button more than five times to move the dot towards the shell) and leave the bad shells (no action was required) according to the probabilistic feedback. After 60 trials, instrumental training ended once participants reached the learning criterion (80% correct choices over the last 16

consecutive trials) or at a maximum of 120 trials. (B) Pavlovian Phase: Participants passively learned the association between five types of compound conditional stimuli (CSs, consisting of fractal-like images and pure tones) and positive (€1, €2), negative (€-1, €-2) or neutral (€0) unconditioned stimuli (USs). There were 80 trials in total with 16 trials of each type. (C) PIT phase: Participants performed the instrumental task again with the tiled fractal images of the CSs in the background. Each trial lasted 3 s, with the fractal images shown 0.6 s before the instrumental shells. Therefore, participants had a response window of 2.4 s. There were 90 trials in total. This phase was done under nominal extinction to avoid further learning. Additionally, there were 72 trials with alcohol/water pictures presented in the background in combination with the two instrumental stimuli (details about the alcohol/water PIT trials shown in Appendix B.6). (D) Query Trials: in order to verify the acquisition of Pavlovian expectations, participants needed to make forced choices between two CSs within 2 s. Each possible pair of the CSs was presented three times in a randomized order.

*Table 2: Drinking behavior of the sample*

| | High-risk drinkers | | | Low-risk drinkers | | |
|---|---|---|---|---|---|---|
| | *N* | Min-max | Mean (SD) | *N* | Min-max | Mean (SD) |
| General description of the sample | | | | | | |
| Age | 94 | 18.1-18.9 | 18.4 (0.2) | 97 | 18.1-18.8 | 18.4 (0.2) |
| Years of Education | 94 | 11-13.5 | 11.6 (0.6) | 96 | 4-14.5 | 11.6 (1.1) |
| Drinking behavior | | | | | | |
| Age 1st drinking | 94 | 10-16 | 14.1 (1.4) | 97 | 9-18 | 14.4 (1.3) |
| Age 1st Drunk | 94 | 12-18 | 15.5 (1.1) | 89 | 10-18 | 16.0 (1.1) |
| Alcohol consumption last year (g/day) | 94 | 3.2-112.5 | 19.4 (16.8) | 97 | 0.6-22.5 | 5.1 (4.6) |
| Alcohol consumption (g/occasion) | 94 | 63-225 | 104.2 (40.4) | 97 | 18-54 | 39.2 (11.5) |
| Age 1st Bingeing | 86 | 14-18 | 16.5 (0.8) | 52 | 14-18 | 16.5 (0.9) |
| Frequency bingeing (lifetime) | 86 | 1-150 | 26.1 (29.7) | 97 | 0-100 | 5.3 (14.3) |
| Alcohol consumption per bingeing (g/occasion) | 94 | 63-450 | 130.9 (52.5) | 97 | 0-225 | 57.2 (59.5) |
| Generic Drink Score* | 94 | -4.5-19.2 | 3.0 (4.2) | 97 | -8.4-8.5 | -2.8 (3.2) |

* Detailed information about how the Generic Drink Score was computed and the statistical analysis regarding this variable are shown in Appendix B.3.

### 3.3.2 Behavioral Analysis

It is important to note that the same dataset was used in a previous study from within our group (Garbusow et al., 2019); however, the analysis of the current study uses these data for a different purpose: to investigate the interference of Pavlovian cues on the ongoing instrumental behavior. A detailed discussion about the difference between the analyses of the current study and Garbusow et al. (2019) is provided in Appendix B.7.

The analysis for this study was restricted to PIT trials that could either be categorized as "congruent" and "incongruent". In the congruent condition, the Pavlovian background value and the instrumental stimulus were positively or negatively concordant, meaning the Pavlovian fractal images corresponding to the monetary gains of 1 or 2 € were paired with the "good" shells. Additionally, the congruent condition consisted of trials in which the Pavlovian images corresponding to monetary losses of 1 or 2 € were paired with the "bad" shells. For the incongruent condition, the opposite is true; this condition consisted of trials that were paired discordantly. To keep the analysis parsimonious, trials with neutral Pavlovian stimuli in the background were disregarded for the analysis. Moreover, trials with or alcoholic/water beverages in the background were also disregarded because it is not clear how healthy young adults would perceive the valence of these backgrounds. Thus, classifying these trials *a priori* as either congruent or incongruent would not have been viable.

The behavioral data were analyzed with R 3.4.0 (R Core Team, Vienna, Austria). ER was used as a primary measurement of task performance in the PIT phase. Correct responses were defined as at least five button presses in collect trials, or less than five button presses in the leave condition, regardless of the background stimuli.

To check whether our approach for PIT data analysis is suitable, we first compared the ER across the 14 conditions (7 Pavlovian cues × 2 instrumental behaviors), which confirmed that the main difference in ER arises from the incongruent versus congruent contrast (Figure S4). Within the incongruent condition, the ER showed a symmetric pattern: collecting a good shell with a negative Pavlovian background did not differ from leaving a bad shell with a positive background. This symmetric pattern held true when assessing the association between the ER and the drinking behavior; a detailed description along with the exploratory analyses of alcoholic/water beverage background trials are displayed in Appendix B.1.

The interference PIT effect score was calculated by subtracting the ER in the congruent condition from the ER in the incongruent condition for each individual. To test whether the participants make more errors in the incongruent condition compared with the congruent condition, a one-tailed, one-sample *t* test was conducted on the interference PIT effect score. The one-tailed test was used on the basis of our *a priori* hypothesis that the ER is higher in the incongruent compared with the congruent condition.

The association between performance during the PIT task and the alcohol drinking behavior was then tested, particularly binge drinking behavior. Again, on the basis of our hypothesis, a one-tailed two-sample *t* test was performed accordingly to test whether the interference PIT effect was higher in the high-risk compared with the low-risk drinking group.

### 3.3.3 fMRI Data Acquisition and Analysis

#### *3.3.3.1 fMRI Data Acquisition and Preprocessing*

The imaging data (Echo-planar imaging [EPI]) sequence and structural T1 weighted image) were acquired using a Siemens 3-Tesla MRI scanner (Magnetom Trio, Siemens, Erlangen, Germany). Preprocessing of the fMRI data was performed with Nipype (Gorgolewski et al., 2011). The 480 EPI images were slice time corrected, realigned to the first image of the sequence, coregistered to the individual segmented and normalized structural image and then smoothed with a Gaussian Kernel of full width at half maximum of 8 mm (see Appendix B.4 for detailed information).

After the preprocessing, 139 subjects were included in the fMRI analysis. Among the 52 subjects who were excluded from the fMRI analysis, there were four participants with incidental findings, 22 participants with more than 3 mm volume-to-volume movement or 3° rotation and 26 participants without valid data for first-level model as they did not press a button at least once for some stimuli, thus having an empty regressor in the first-level model preventing model estimation.

#### *3.3.3.2 fMRI Data Analysis*

Statistical analyses of the fMRI data during the PIT phase were performed by the general linear model (GLM) in SPM12 (Wellcome Trust Centre for Neuroimaging, London, UK). On the first level, a model that consisted of 10 onset regressors of our main interest was used: five

Pavlovian CS values (€-2, -1, 0, 1, and 2 monetary loss or reward) × two instrumental conditions (collect or leave). Additionally, four onset regressors for the alcohol/water trials (collect/leave × alcohol/water) were also included in the first-level model. The onset of each registered button press was entered into a regressor of no interest. Finally, six nuisance (motion) regressors were also included in the model.

On the first level, the incongruent versus congruent contrast was defined as follows: the four types of incongruent trials were collapsed (CSs paired with €-1 or -2 in the collect trials or CSs paired with €1 or 2 in the leave trials) together and then the four types of congruent trials were subtracted, thus mirroring the behavioral analysis. These individual contrast images were then entered into second-level SPM analysis (one-sample *t* test). To associate the neural incongruency effect (i.e., brain response to interference) with the behavioral performance at the group level, the individual behavioral interference PIT effect was included as a covariate in the second-level model. Additionally, a covariate of no interest was also included to specify the site information (the experiment was performed in either Berlin or Dresden). To test the hypotheses, brain responses in four regions of interest (ROIs) were analyzed. The dmPFC, lPFC, and VS masks were defined on the basis of the 12 mm spheres around the peaks from previous review papers (Hung et al., 2018; Liu et al., 2011). The amygdala mask was defined anatomically (details in Figure 8). The mean individual parameter estimates were then extracted within the four ROIs from the first-level incongruent versus congruent contrast. To examine the neural incongruency effect on the group level, the mean parameter estimates from the four ROIs were tested in 4 one-sample *t* tests. Following this, the association between the brain response to interference and the behavioral interference PIT effect ($\Delta$ ER) was tested with Pearson correlation tests for the four ROIs separately. These results were corrected for four comparisons with Bonferroni correction, with $p_{corr.} = .05$ ($p_{uncorr.} = .0125$) as the threshold.

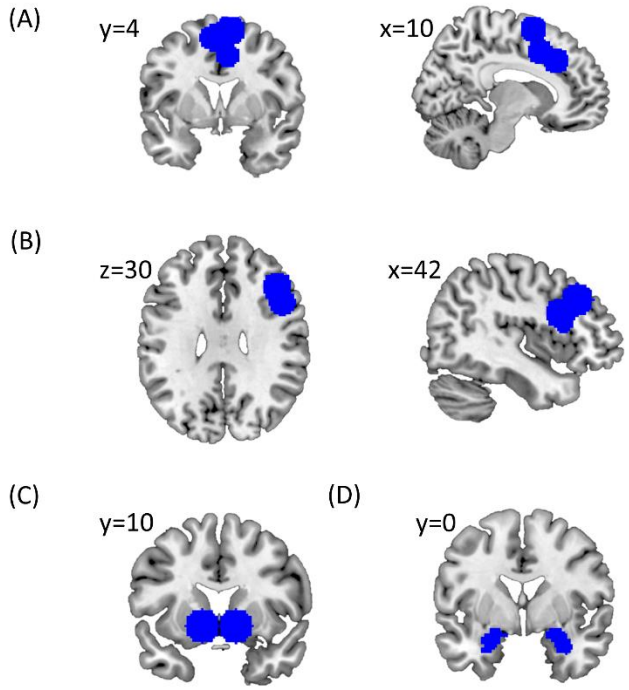


**Figure 8: Regions of interest (ROI) masks.** (A) dorsomedial prefrontal cortex (dmPFC) mask: generated according to the review paper by Hung et al. (2018). In the cognitive inhibition meta-analysis, there were four peaks located in the dmPFC regions (Talairach Coordinate: 6/14/40; 6/26/32; 8/8/58; -6/0/54). Four 12 mm spheres were generated around each peak, and the conjunctions of these spheres were used as the dmPFC masks. (B) lateral prefrontal cortex (lPFC) mask: conjunction of the three 12 mm spheres generated around the three peaks in the lPFC according to the same meta-analysis (Talairach coordinates: 42/26/30; 46/14/22; 52/16/14) (C) ventral striatum (VS) mask: defined based on the peak of a previous meta-analysis on functional magnetic resonance imaging (fMRI) reward-related tasks (Liu et al., 2011). The conjunction of the two 12 mm spheres around the peak MNI coordinates: -12/10/-6 and 12/10/-6 were defined as the VS mask. (D) Amygdala mask: the bilateral amygdala mask was defined anatomically on the basis of the AAL atlas in the WFU PickAtlas toolbox (Tzourio-Mazoyer et al., 2002).

These ROI analyses were followed by an exploratory whole-brain analysis of the incongruent versus congruent contrast, as well as its association with the behavioral interference PIT effect (i.e., covariate effect on the second level) at an uncorrected threshold of $p < .001$, cluster size $k \geq 50$. Whether or not the association between behavioral and neural incongruency effect differs from risk status was also explored. The detailed description for this analysis is shown in B.5 in the Appendix.

To further explore how the effective connectivity modulated by the incongruent condition differs between the two groups, especially regarding the interplay between the VS and the dmPFC and lPFC areas, dynamic causal modelling (DCM) analyses were applied to the data (Friston et al., 2003). The time series were extracted from the peak voxels within the VS, lPFC and dmPFC that showed more activation during the conflict (i.e., incongruent-congruent contrast) because no regions were less activated during the conflict. Accordingly, for each individual, the time series of the three regions were extracted from 8 mm spheres centered on the individual local maxima, which were allowed to vary within 5 mm spheres around the group peak voxels during the conflict (incongruent-congruent contrast). The amygdala was excluded for this exploratory analysis, as there was no neural response in the amygdala within our contrast of interest; detailed information about this can be found in Section 3.4. In the model space, full intrinsic connections were assumed among the three regions, including self-connections. All PIT trials were used as driving inputs to enter VS, and the incongruent condition was used as modulatory input. Three possible modulatory effects were assumed on the connections between each pair of the three regions: forward, backward or bidirectional. With three possible connection structures between each pair, our model space consisted of 27 models in total (three possible structures × three pairs between the three regions; Figure 9).

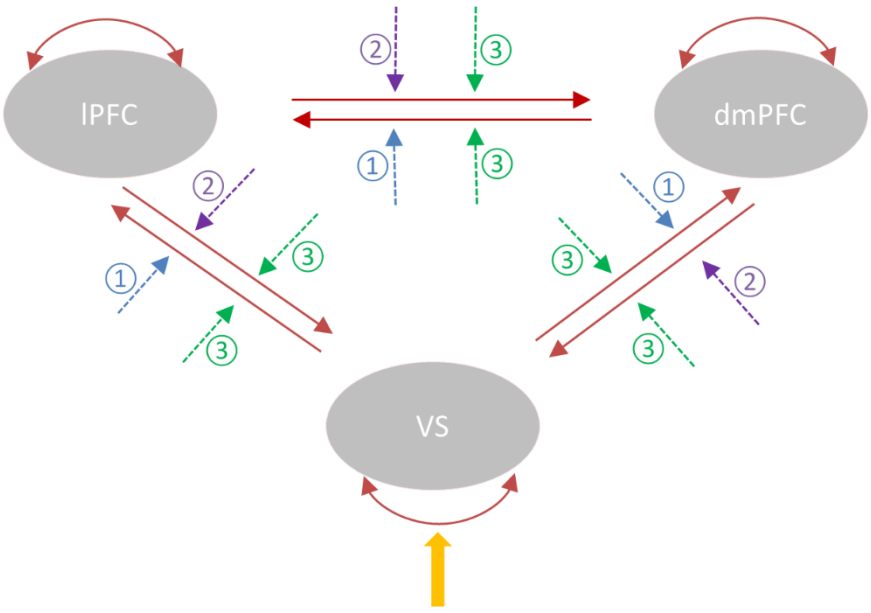


**Figure 9 Dynamic causal modelling (DCM) model space.** There were 27 models in the DCM model space. The driving input consisted of all Pavlovian-to-instrumental transfer (PIT) trials that entered the

ventral striatum. The red arrows specify the intrinsic connectivity: all three regions were assumed to be intrinsically connected to each other and to themselves. The incongruent condition was assumed to modulate the connectivity between each pair of regions in three ways: forward, backward, or bidirectional, which resulted in 27 modulatory structures in total.

Following this, Bayesian model selection (BMS) was conducted in combination with family-level inference (Penny et al., 2010). The aim of the family-level inference, in this case, was to compare the models with different types of interplay between the VS and the two prefrontal areas during the incongruent condition. Six families (Figure 10) were defined accordingly: (1) incongruent condition only modulates the top-down connections; (2) incongruent condition only modulates the bottom-up connections; (3) incongruent condition modulates top-down connections between the VS and the dmPFC but bottom-up connections between the VS and the lPFC, or vice versa; (4) incongruent condition modulates both top-down and bottom-up connections only for the lPFC; (5) incongruent condition modulates both top-down and bottom-up connections only for the dmPFC; (6) incongruent condition modulates both top-down and bottom-up connections for both the lPFC and the dmPFC. The BMS was done separately for the two groups. Given that fixed optimal model structures were not assumed among individuals, a random-effects analysis was used on the group level. This method takes into account the individual differences in model structures (Stephan et al., 2010). Following the BMS, Bayesian model averaging (BMA) was performed across the entire model space to further obtain parameter estimates of the effective connectivity. Finally, two-sample *t* tests were done to compare the connectivity between the two groups. The results were corrected for six comparisons with Bonferroni correction, with $p_{corr.} = .05$ as the threshold.

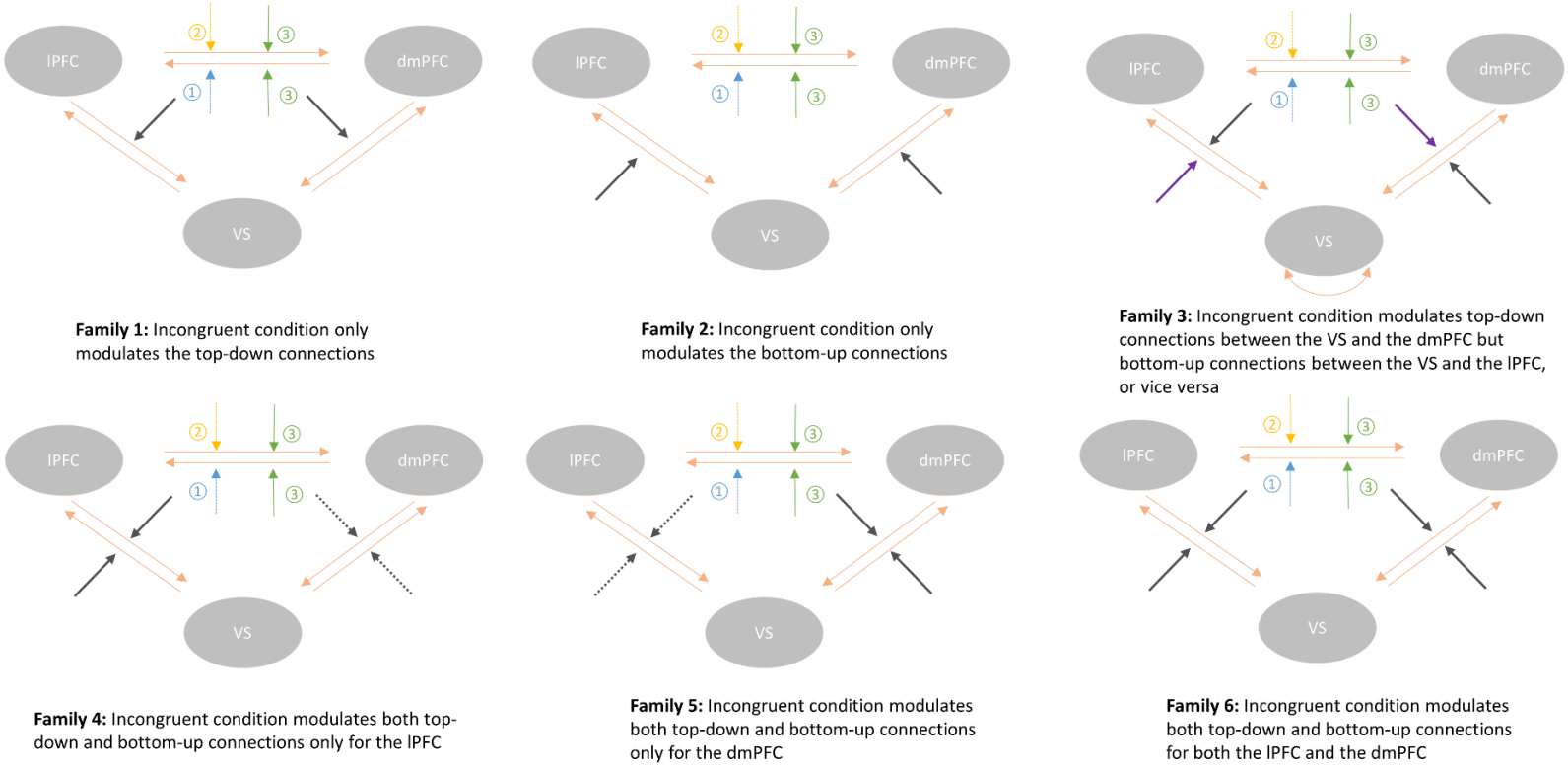


**Figure 10: Dynamic causal modelling (DCM) model families.** The 27 DCM models were divided into six model families on the basis of the modulatory effect of the incongruent condition on the connectivity between the ventral striatum (VS) and the two prefrontal regions. Within each model family, there were three possible types of modulatory effects of the incongruent condition on the lateral prefrontal cortex (lPFC) – dorsomedial prefrontal cortex (dmPFC) connection: forward, backward, and bidirectional.

### 3.3.4 Association Between Risk Status and PIT Effect

To further examine whether the PIT effects were associated with risk status, logistic regression was employed with risk status as the dependent variable. Possible predictors included the behavioral interference PIT effect and parameter estimates from the neural activated clusters in the incongruent condition (after adjusting for the behavioral interference PIT effect to avoid collinearity in predicting). In a stepwise backward selection process, the best combination of predictors was examined. Data-driven clusters were again used for this analysis, because it was expected that these regions would reflect the neural responses within the PIT task more precisely compared with the ROIs.

## 3.4 Results

### 3.4.1 Behavioral Results

The ER was found to be, on average, approximately twice as high in the incongruent condition (30.8%) as compared with the congruent condition (15.6%, Figure 11A). This increase of ER was highly significant ($T$ = 7.23; df = 190; $p = 5.47\times10^{-12}$; d = 0.52), indicating a substantial interference PIT effect in the whole sample. As hypothesized, the PIT effect was substantially stronger in the high-risk compared to the low-risk drinking group ( $\Delta$ ER $_{\text{high-risk}}$ = 21.3%, $\Delta$ ER $_{\text{low-risk}}$ = 9.2%, $T$ = 2.96; df = 189; $p = 1.74\times10^{-3}$; d = 0.43). The results are displayed in Figure 11B. *t* tests on working memory, processing speed and crystallized intelligence revealed no significant differences between the two groups (for details, see Appendix B.2).

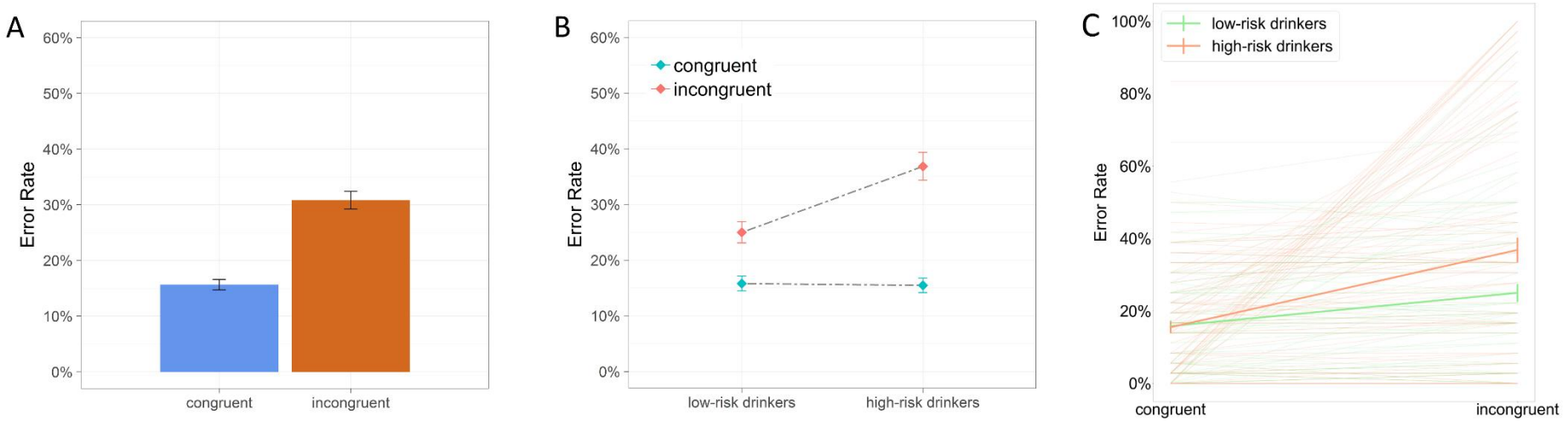


**Figure 11: Behavioral interference Pavlovian-to-instrumental transfer (PIT) effect.** (A) Error rate (ER) increased on average by 15.2% in the incongruent condition compared with the congruent condition ($p$ = 5.47×10$^{-12}$). (B) High-risk drinkers ($N$ = 94), in contrast to the low-risk drinkers ($N$ = 97), reflected increased ER in the incongruent condition compared with congruent condition ($p$ = 1.74×10$^{-3}$). (C) Individual ER change in the incongruent compared with the congruent condition, separated between high- and low-risk drinkers.

### 3.4.2 fMRI Results

*3.4.2.1 Neural Incongruency Effect—ROI Analysis*

In the ROI analyses, the four one-sample *t* tests of the parameter estimates within the four ROIs did not survive the correction for multiple comparisons, thus indicating no significant difference in the congruent condition compared with the incongruent condition on the group level.

*3.4.2.2 Neural Correlates of the Behavioral Interference PIT Effect—ROI Analysis*

When correlating the behavioral interference PIT effect and neural responses (incongruent - congruent condition) in the four ROIs, positive correlations were found between the behavioral interference PIT effect ($\Delta$ ER) and the neural responses in the lPFC ($r(137) = 0.23$; $p_{\text{one-tailed; corr.}} = .012$) as well as in the dmPFC ($r(137) = 0.25$; $p_{\text{one-tailed; corr.}} = .007$). The correlation between neural responses in the VS and the behavioral interference PIT effect was also positive, but it did not survive the control for multiple comparisons ($r(137) = 0.16$; $p_{\text{one-tailed}} = .080$ without the Bonferroni correction). However, correlations were not seen between the behavioral interference PIT effect and responses in the amygdala ($r(137) = -0.02$; $p_{\text{one-tailed}} = .790$ without the Bonferroni correction).

*3.4.2.3 Neural Incongruency Effect—Whole-Brain Analysis*

With respect to the explorative whole-brain analysis, the second-level *t*-contrast of the incongruent versus congruent PIT condition was first investigated; this included the individual behavioral interference PIT effect as a covariate. Increased brain responses during the incongruent compared with the congruent PIT trials (neural incongruency effect) were found in the ventral tegmental area (VTA; $k = 50$, $T = 4.01$, peak MNI (Montreal Neurological Institute [MNI] templates) coordinates: -10/-16/-22) at a whole-brain uncorrected threshold of $p < .001$, cluster size $k \geq 50$ (Figure 12A). As an additional sanity check, at a lower threshold ($p < .01$, cluster size $k \geq 50$), the BOLD response of parietal top-down control regions (BA40, peak MNI coordinates: -34/-48/50, $k = 265$, $T = 2.93$) were also more pronounced during the incongruent condition. In contrast, no brain region showed higher activity during the congruent compared with the incongruent PIT trials at the same statistical threshold (whole-brain $p < .001$, cluster size $k \geq 50$).

*3.4.2.4 Neural Correlates of the Behavioral Interference PIT Effect—Whole-Brain Analysis*

In the next step of the whole-brain analyses, whether or not the neural response to interference was associated with the behavioral interference PIT effect was investigated by conducting a one-sample *t* test on the behavioral interference PIT effect covariate. Neural correlates of the behavioral interference PIT effect were seen in the VS ($k$ = 168, $T$ = 4.58, peak MNI coordinate: 14/16/0), lPFC ($k$ = 235, $T$ = 3.97, peak MNI coordinate: 50/38/22), and dmPFC ($k$ = 955, $T$ = 4.35, peak MNI coordinate: 8/20/48) at a whole-brain uncorrected threshold of $p < .001$, $k \geq 50$ (Figure 12B; detailed results displayed in Table 3). To illustrate the brain correlates of the behavioral interference PIT effect ($\Delta$ER), the neural activation within the three activated clusters was plotted in response to incongruent over congruent trials (neural incongruency effect) against the behavioral interference PIT effect (Figure 13). As can be seen, the neural response to incongruency in the VS, lPFC and dmPFC was higher in subjects with a stronger behavioral interference PIT effect. However, not all the individuals showed responses to incongruency—this effect was driven by around half of the individuals who committed more errors in the incongruent condition as compared with the congruent condition. The association between the behavioral interference PIT effect and the neural incongruency effect was stronger for low-risk drinkers compared to high-risk drinkers in the VS and the lPFC, but the difference was marginal in the dmPFC (detailed result in Figure S6 and Figure S7 in the Appendix).

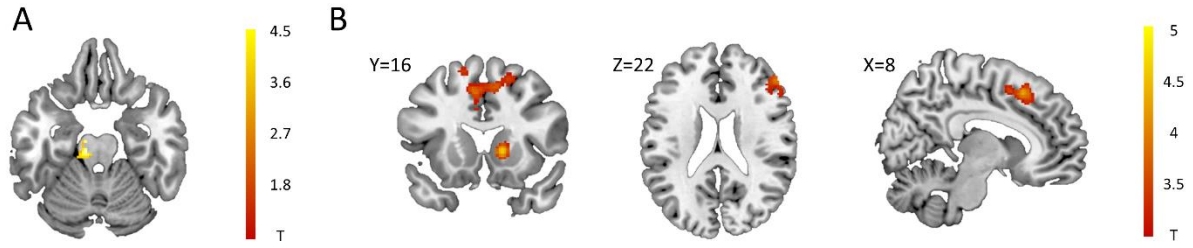


**Figure 12: Neural incongruency effect & neural correlates of behavioral interference Pavlovian-to-instrumental transfer (PIT) effect** ($p_{uncorrected} < 0.001$, cluster size $k \geq 50$). **(**A). Interference (incongruent – congruent trials) elicited activation in the ventral tegmental areas (VTA) ($T$ = 4.01, $k$ = 50, peak Montreal Neurological Institute [MNI] coordinates: -10/-16/-22). (B) A Neural PIT effect (brain response to interference correlated with behavioral interference PIT effect) was found in the ventral striatum (VS) ($T$ = 4.58, $k$ = 168, peak MNI coordinates: 14/16/0), lateral prefrontal cortex (lPFC) ($T$ = 3.97, $k$ = 235, peak MNI coordinates: 50/38/22) and dorsomedial prefrontal cortex (dmPFC) ($T$ = 4.35, $k$ = 955, peak MNI coordinates: 8/20/48).

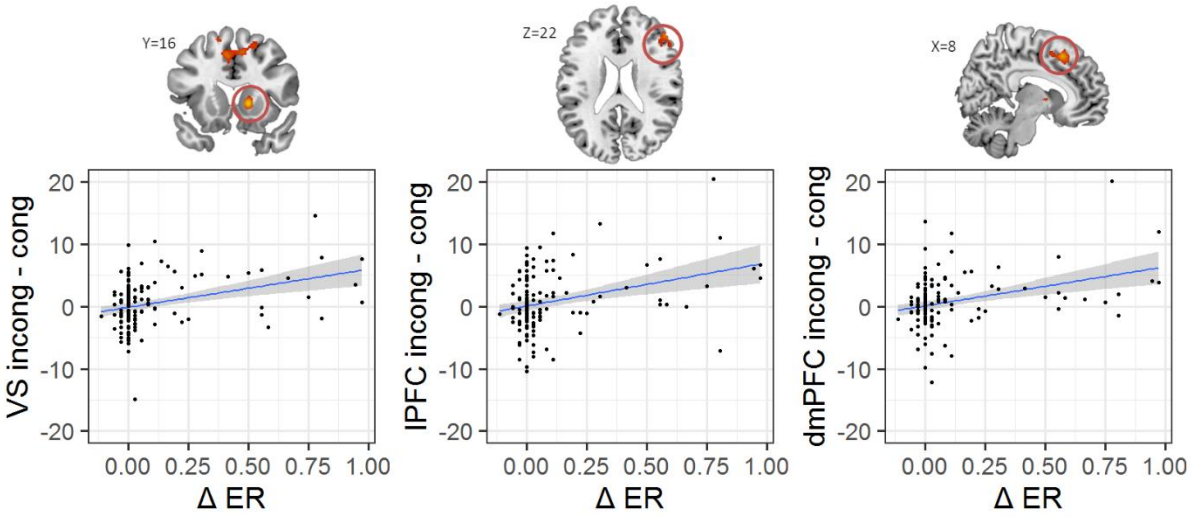


**Figure 13: Neural correlates of behavioral interference Pavlovian-to-instrumental transfer (PIT) effect.** Illustration of the positive association between neural activation in the ventral striatum (VS), lateral prefrontal cortex (lPFC), dorsomedial prefrontal cortex (dmPFC) and the behavioral interference PIT effect.

*Table 3: fMRI results table*

| Whole-brain results ($p_{uncorrected.}$ < .001, cluster size ≥ 50) | | | | | | |
|---|---|---|---|---|---|---|
| Region | Side | Peak MNI | | | Peak -level *t* score | Cluster size |
| | | *x* | *y* | *z* | | |
| **Neural incongruency effect (incongruent – congruent)** | | | | | | |
| Brain-stem (midbrain) | L | -10 | -22 | -22 | 4.19 | 50 |
| Inferior temporal gyrus | R | 58 | -42 | -16 | 3.76 | 157 |
| **Neural activation in association with the behavioral PIT effect** | | | | | | |
| Right ventral striatum (extended to caudate) | R | 14 | 16 | 0 | 4.58 | 168 |
| SMA (BA32, extended to BA8 and BA6) | R | 8 | 20 | 48 | 4.35 | 955 |
| Middle frontal gyrus (SMA; BA 6) | L | -28 | 2 | 58 | 4.03 | 226 |
| Middle frontal gyrus (DLPFC/VLPFC; BA 45) | R | 50 | 38 | 22 | 3.97 | 235 |
| Middle frontal gyrus (IFG; BA 44) | L | -36 | 22 | 34 | 3.84 | 69 |

Abbreviations: DLPFC, dorsal lateral prefrontal cortex; fMRI, functional magnetic resonance imaging; IFG, inferior frontal gyrus; MNI, Montreal Neurological Institute; PIT, Pavlovian-to-instrumental transfer; SMA, supplementary motor area; VLPFC, ventral lateral prefrontal cortex.

#### *3.4.2.5 Effective Connectivity Difference Between High- and Low-Risk Drinkers*

The model selection was first performed in order to select an optimal family of models among the six families in Figure 10. The selection was performed separately for the high- and low-risk drinking groups to test whether the winning family of models was different for the two groups. The selection was based on the exceedance probability: a higher exceedance probability suggests one family of models has more evidence compared with other specified families of models. According to the family exceedance probability, the winning family for the high-risk drinking group was the family in which the incongruent condition only modulated the bottom-up but not the top-down connections between the VS, lPFC, or the dmPFC. The winning family had an exceedance probability of 0.32 (compared to the second-best family with an exceedance probability of 0.17). In contrast, for the low-risk drinkers, the model family in which the incongruent condition fully modulated all the connections between the

VS and both the lPFC and the dmPFC had the highest exceedance probability of 0.38 (the second-best family had an exceedance probability of 0.19). Generally speaking, with around twice the exceedance probability of the winning family compared to the second-best family, it was concluded that there was only weak support for the two different winning families for the two groups (plotted in Figure 14). Because of the different winning families, the strength of the connectivity was further obtained through BMA across the entire model space for both groups; this ensured the parameter estimates were comparable. The BMA does not make inferences about the model structure, but it rather computes a weighted average of the effective connectivity parameters from all the specified models. The weights are given by the posterior probabilities of different models (Stephan et al., 2010). On the basis of the BMA results, one can directly compare whether the effective connectivity parameters between certain brain regions are different for the two groups. According to the criteria that the posterior mean is larger than zero at a probability threshold of 95%, the incongruent condition significantly modulated the connection from the VS to the lPFC and the bidirectional connection between the lPFC and the dmPFC for the low-risk but not the high-risk drinkers (Table 4). By comparing the modulatory parameters between the two groups, significantly higher effective connectivity was found from the VS to the lPFC modulated by the incongruent condition in the low-risk compared to the high-risk drinking group ($p$ = .004 after Bonferroni correction for six comparisons) (Table 4).

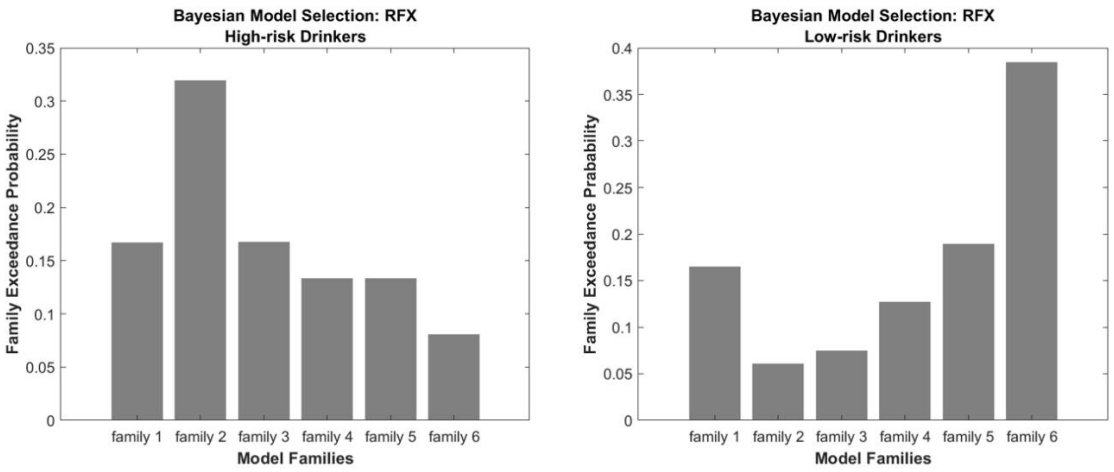

**Figure 14: Bayesian model selection (random-effects analysis; RFX) results for the high-risk and low-risk drinkers.** According to the family exceedance probability, the winning family for the high-risk drinking group was the family where incongruent condition only modulates the bottom-up but not the top-down connections between the ventral striatum (VS) and the lateral prefrontal cortex (lPFC) as well as the dorsomedial prefrontal cortex (dmPFC) (Family 2). In contrast, for the low-risk drinkers, the model family where the incongruent condition fully modulates all the connections between the VS and both the lPFC and dmPFC had the highest exceedance probability (Family 6).

*Table 4: DCM results*

| **Modulatory effects of the incongruent condition** | | | | |
|---|---|---|---|---|
| | Low-risk drinkers | High-risk drinkers | **Two-sample *t* test** | |
| | | | *t* value | *p* value |
| VS→lPFC | 0.056 (.099) ** | -0.002 (.095) | 3.52 | .001 ** |
| VS→dmPFC | 0.021 (.097) | 0.017 (.093) | 0.22 | .829 |
| lPFC→VS | 0.001 (.098) | 0.004 (.097) | -0.22 | .828 |
| lPFC→dmPFC | 0.049 (.100) ** | 0.013 (.097) | 2.13 | .035 * |
| dmPFC→VS | -0.010 (.098) | 0.006 (.097) | -1.00 | .317 |
| dmPFC→lPFC | 0.045 (.099) ** | 0.020 (.097) | 1.52 | .132 |
| **Driving input from all PIT trials** | | | | |
| → VS | 0.011 (.008) ** | 0.005 (.009) | 3.98 | .001 ** |

Abbreviations: DCM, dynamic causal modelling; dmPFC, dorsomedial prefrontal cortex; lPFC, lateral prefrontal cortex; PIT, Pavlovian-to-instrumental transfer; VS, ventral striatum.

* Significant at uncorrected threshold $p < .05$

** Survives Bonferroni correction for multiple comparisons (six comparisons)

### 3.4.3 Association Between Risk Status and PIT Effects

In the backward stepwise logistic regression with risk status as the dependent variable, the best model ($\chi2(3, N = 139) = 8.966$, $p = .030$) included three of the four predictors: the behavioral interference PIT effect ($\beta = 2.073$; $p = .014$), the neural activation in the incongruent condition in the VS ($\beta = 0.298$; $p = .091$) and the lPFC ($\beta = -0.391$; $p = .042$), but not in the dmPFC. The logistic regression thus indicated a positive association between risk

status and behavioral interference PIT effect and the VS (trend-wise), whereas the risk status was negatively associated with the neural responses in the lPFC.

## 3.5 Discussion

In this study, we investigated whether interference between Pavlovian and instrumental control, assessed with a PIT task, is associated with risky alcohol use in a cohort of healthy males aged 18 years. As expected, participants committed substantially more errors in the incongruent compared with the congruent condition, which suggests that interference by incongruent Pavlovian cues impairs instrumental performance. Importantly, the instrumental performance was substantially more impaired by Pavlovian interference in high-risk compared with low-risk drinkers, indicating better interference control abilities in the latter. At the neural level, participants with a stronger behavioral instrumental impairment showed higher activation in the VS, the dmPFC, and the lPFC during incongruent PIT trials. Furthermore, the neural responses also differed with risk status: high-risk drinkers showed a blunted top-down control response of the lPFC, as well as reduced effective connectivity from the VS to the lPFC during the incongruent (i.e., conflict) condition. Taken together, these findings indicate that individuals who can allocate top-down control to overcome conflict, that is, interference between Pavlovian and instrumental cues, are less likely to show risky alcohol consumption.

At the behavioral level, the effect of interference was very pronounced; however, at the neural level, interference was not detected in the *a priori* ROIs. The subsequent explorative whole-brain analysis revealed that incongruence was reflected by stronger activation in the VTA and parietal areas, but these activations would not have survived correction for multiple comparisons. Thus, for the entire sample of young males, the neural effect of interference between Pavlovian and instrumental control was rather modest. Regarding brain regions, this finding is in line with previous animal studies, which showed that inactivation of the VTA reduced the PIT effect (Corbit et al., 2007; Murschall & Hauber, 2006). Additionally, activation of the parietal areas, which has been suggested to be part of the inhibitory brain network (Hung et al., 2018), may indicate the conflict participants experienced in the incongruent condition. The modest effect on the group level might be due to the fact that only about half

of the sample showed impaired performance during interference between instrumental and Pavlovian control.

In contrast, when the interindividual differences in interference were considered, it was found that the VS, lPFC and dmPFC activation correlated positively with the behavioral interference PIT effect. Previous literature repeatedly reported the VS to reflect the influence of the Pavlovian cue on instrumental behavior (Geurts et al., 2013; Mendelsohn et al., 2014; Talmi et al., 2008). The VS cluster that was found also extended to the dorsal striatum; this has also been shown by two previous studies (Bray et al., 2008; Lewis et al., 2013). In contrast to previous studies, we did not find amygdala activation (Garbusow et al., 2019; Geurts et al., 2013; Mendelsohn et al., 2014; Prevost et al., 2012; Talmi et al., 2008). As suggested by these studies, the amygdala may compute the affective valence of Pavlovian cues in the PIT task. Notably, one difference between the previously mentioned studies and the current study involves the valence signal. In the aforementioned PIT studies, when comparing the positive/negative Pavlovian cue condition with the neutral condition, the finding reflected a mixture of salience and valence signal. Conversely, in the current analysis, the valence signal was averaged out when pooling the different combinations of Pavlovian cues and instrumental stimuli into incongruent and congruent conditions. This may begin to explain why activation in the amygdala was not found. Taken together, the signal seen in the VS may reflect a salience signal indicating that the Pavlovian cue is at odds with the required instrumental behavior.

The response elicited by incongruent trials was also found in the dmPFC. This region has been extensively linked to conflict-related performance monitoring, in which it plays an important role in deciding the subsequent adjustments in performance (Domenech & Koechlin, 2015; Ridderinkhof et al., 2004). Additionally, incongruent trials also evoked a response of the lPFC, which is a critical structure that gathers task-related information and exhibits top-down cognitive control (Egner & Hirsch, 2005; Kouneiher et al., 2009) in relation to conflict monitoring, error monitoring and response selection (Amodio & Frith, 2006). To summarize, the activation found in the VS, lPFC and dmPFC is part of a corticostriatal circuit that is critical for response selection and cognitive control through the extensive communication between the subcortical and cortical parts (Haber, 2016; Peters et al., 2016)—which makes it essential for overcoming interference during incongruent task trials.

Compared with low-risk drinkers, the high-risk drinkers showed a stronger association between the behavioral and the neural PIT effect. This effect may be related to the findings from the DCM analysis, which suggested that the incongruent stimuli tended not to modulate the effective connectivity from the dmPFC and lPFC to the VS for the high-risk drinkers. Parameter estimates further indicated that the effective connectivity from the VS to the lPFC was higher in response to the incongruent stimuli in the low-risk compared to the high-risk drinking group. It is also worth mentioning that the VS mask for the DCM analysis was generated around the peak activation from the analysis—this mask also partly consisted of the dorsal striatum. Therefore, the interplay between the VS and the lPFC may have also involved the dorsal striatum to some extent. Taken together, the neural response in this network may explain why low-risk drinkers showed better interference control (i.e., were less susceptible to response conflicts induced by incongruent stimuli) when the Pavlovian cue conflicts with the instrumental behavior. It is plausible that the VS of low-risk drinkers sends a salience signal that helps allocate cognitive top-down control to resolve the response conflict.

It is worth noting that a previous paper from our group found that the association between the valence of the Pavlovian cues and response rates (indicating response vigor) was stronger for high-risk than low-risk drinkers (Garbusow et al., 2019). However, in this study, the main focus was to investigate the motivational effect of Pavlovian cues on the ongoing instrumental behaviors, regardless of whether they promote (congruent condition) or hinder (incongruent condition) the required instrumental response. Despite using the same dataset, the main focus of the current study was to examine the interference effect of Pavlovian cues when they are in conflict with the necessary instrumental behavior. By doing this, the motivational and cognitive control perspectives were able to be examined simultaneously, as both perspectives were present during trials with interference from Pavlovian cues. Therefore, these results connect previous research in the fields of cognitive control and motivated behavior. Even though the interplay of cognitive control and motivated behavior is essential to understand addictive behavior, most experimental approaches either focus on one or the other. An exception would be the go-no-go/PIT task (Freeman et al., 2015; Freeman et al., 2014), which assesses the influence of non-drug Pavlovian cues on response inhibition. So far, go-no-go/PIT tasks have not been used to study substance use or dependence. These results, therefore,

complement previous studies that reported an association between binge drinking and impaired interference control in young adults (Carbia et al., 2018).

Importantly, the conflict between Pavlovian and instrumental control substantially differs from conflict seen in traditional interference tasks such as the classical color-word Stroop task (conflict at stimulus level) (Macleod, 1992; Stroop, 1935) or the Simon task (conflict at response level) (Hommel, 2011; Simon & Rudell, 1967). In these "cold" interference tasks, responses are instructed and are not the result of learning based on rewards or punishments. Interference in these tasks mainly results from automated response tendencies (i.e., neither the color representation in the Stroop task nor the location cue representation in the Simon task triggers motivational responses). In contrast, in our "hot" interference task, Pavlovian cues trigger a motivational response, that is, approach or avoidance behavior and interfere with motivated instrumental behavior. On the basis of the hypothesis about the difference between the "cold" and "hot" interference task, future studies could investigate whether the PIT effect we found could (to some extent) be explained by these "cold" interference tasks or it involves fundamentally different mechanisms.

To conclude, the results of the current study show that the susceptibility to Pavlovian interference during a PIT task is linked to hazardous drinking behaviors at age 18. Although the imbalance between the top-down and bottom-up systems has been suggested to be associated with addictive behavior, previous studies have tended to consider either the perspective of cognitive control or motivated behavior, but not both at the same time. Using a PIT task, we assessed the top-down control and its interaction with bottom-up Pavlovian and instrumental processes. Our experimental data indicate that a poor interplay between top-down and bottom-up processes may contribute to early hazardous alcohol use.

## 3.6 Limitation

We investigated a sample of 18-year-old social drinkers. In this sample, some participants did not commit any errors during the PIT task. It is thus unclear whether these participants experienced no interference at all or they had better interference control. Another explanation could be that the PIT task was not sensitive enough to capture the very subtle effects that may have been present in these participants. Therefore, a possible solution to this issue could be found in further refinement of the PIT task to increase the sensitivity to

more subtle effects. Additionally, the classification of high- and low-risk drinkers based on the self-reported alcohol consumption data during the past year may not be entirely accurate because of the possible memory bias; future studies may improve this by using more frequently assessed electronic diary data. Another limitation of the current study is that these results cannot be generalized to non-male populations.

## 3.7 Acknowledgements

This study was supported by the Deutsche Forschungsgemeinschaft (DFG; grants for FOR 1617 [Project number 186318919], TRR 265 [Project number 402170461] (Heinz et al., 2020) and SFB 940 [Project number 178833530]) , SN received funding from the UZH Grants Office (FK-19-020). MR-imaging for this study was in part performed at the Berlin Center for Advanced Neuroimaging (BCAN). We thank Matthew Belanger for his proofreading. Open access funding enabled and organized by Projekt DEAL.

# Chapter 4: PIT and the Six-Year Risky Drinking Trajectory (Study 3)

This chapter is under major revision in *Addiction Biology* as: **Chen, H.**, Belanger, M. J., Garbusow, M., Kuitunen-Paul, S., Huys, Q. J. M., Heinz, A., Rapp, M. A. & Smolka, M. N. Susceptibility to interference between Pavlovian and instrumental control predisposes risky alcohol use developmental trajectory from ages 18 to 24.

## 4.1 Abstract

We recently reported that susceptibility to interference of Pavlovian and instrumental control assessed via a Pavlovian-to-instrumental transfer (PIT) task was associated with risky alcohol use at age 18. Through latent growth curve modelling, we now investigated whether such susceptibility also predicts the drinking trajectories until age 24. The interference effect during PIT, assessed at ages 18 and 21 during fMRI, was behaviorally characterized by an increase in error rate (ER) during conflict, i.e., when a required instrumental action associated with positively-valenced instrumental cue was performed in the presence of a negatively-valenced Pavlovian cue or vice versa. Functional imaging revealed that the interference PIT effect was characterized by neural responses in the ventral striatum (VS) and the lateral and dorsomedial prefrontal cortices (lPFC and dmPFC, respectively). Drinking trajectories were based on the AUDIT-C (Alcohol Use Disorders Identification Test consumption score) and a binge drinking score (gram alcohol / drinking occasion). We found that a stronger VS response during conflict at age 18 was associated with a higher starting point of both drinking trajectories but was negatively associated with the development of the binge drinking score trajectory. At age 21, high ER and enhanced neural responses in the dmPFC were associated with a risky AUDIT-C trajectory that started to emerge and develop until age 24. Overall, the susceptibility to interference between Pavlovian cues and instrumental control could be viewed as a predisposing mechanism towards hazardous alcohol use during young adulthood, and the identified high-risk group may profit from targeted interventions.

## 4.2 Introduction

The interaction of Pavlovian conditioned cues with instrumental behavior may explain how certain stimuli can trigger drug-seeking in spite of conscious decisions against consumption

(Everitt & Robbins, 2016). The Pavlovian-to-instrumental transfer (PIT) paradigm is an essential experimental tool that allows for the investigation of the influence of Pavlovian cues on ongoing instrumental behavior. Previously, we have demonstrated that susceptibility to interference caused by non-drug related Pavlovian cues that conflict with required instrumental behavior is associated with risky drinking behavior at age 18 (Chen et al., 2021d). Notably, decreased functional activation elicited by the PIT effect in the lateral prefrontal cortex and a trend towards increased activation in the ventral striatum was associated with high-risk drinking. These results suggest a tipping of the balance between cortical and subcortical activation during PIT towards the ventral striatum, which may impact on inhibitory control, risk-seeking behavior, and the motivation to consume drugs (Koob & Volkow, 2016). It is thus of interest to assess how the interference effect during the PIT task, on both the behavioral and neural level, is associated with the development of risky drinking behavior during the early intoxication and binge drinking phases in young adults.

Our PIT experiment (Garbusow et al., 2014) is comprised of two phases that separately motivate instrumental and Pavlovian learning with monetary outcomes. Transfer effects are then assessed in a third phase, during which the participant must provide instrumental responses in the presence of Pavlovian cues from part two. Previous research has demonstrated that the valence of the Pavlovian cues could influence instrumental responding. Specifically, appetitive Pavlovian cues could promote approach or inhibit avoidance, while aversive Pavlovian cues could promote avoidance or inhibit approach behavior (Geurts et al., 2013; Huys et al., 2011; Huys et al., 2016). Previous studies from our group detected increased instrumental responding with respect to the Pavlovian-associated monetary outcomes that incrementally increase in value among patients with alcohol dependence, and increased brain activity in the nucleus accumbens among patients with a poor treatment outcome (Garbusow et al., 2016; Garbusow et al., 2014; Schad et al., 2019).

Employing an additional approach to the analysis, which considers both the Pavlovian cue valences and the required (approach or avoidance) instrumental actions, we have identified further differences in instrumental behavior based on the congruity between the two. To elaborate, Pavlovian cues can interfere with a required instrumental response when they are incongruent with the expected outcome. For example, when an approach response is required in the presence of a negatively-valenced Pavlovian cue, the participant may

erroneously provide an “avoid” response. This interference effect of Pavlovian cues on the instrumental behaviors can be assessed by the error rate, which was indeed found to be higher in the incongruent as compared with the congruent condition (Chen et al., 2021d; Sommer et al., 2020; Sommer et al., 2017). Importantly, patients with Alcohol Use Disorder (AUD), particularly future relapsers, were shown to commit more errors in the incongruent condition than control participants (Sommer et al., 2020; Sommer et al., 2017). The same effect was also found in high-risk compared to low-risk drinkers at age 18 in the preclinical group investigated in the current study (Chen et al., 2021d). However, a recent study that assessed a full PIT task using food rewards found that the valence of the Pavlovian cues did not influence the performance of the AUD and the control group differently (van Timmeren et al., 2020).

On the neural level, we have previously shown that a higher error rate during the incongruent condition was associated with stronger neural responses in the ventral striatum (VS) and the lateral and dorsomedial prefrontal cortices (lPFC and dmPFC). This finding suggests relationships between the influence of the Pavlovian cues and ongoing instrumental behavior that encompass both bottom-up and top-down neural pathways. Top-down cognitive control may play a critical role in this relationship, especially when the Pavlovian cues conflict with the instrumental responses. Support for the hypothesis that top-down control plays a key role during conflict trials comes from another school of literature in which a valenced go-no-go task was used. In this task, instead of assessing the PIT effect during a separate transfer phase following the instrumental and Pavlovian trainings, the Pavlovian conflict was embedded in the ongoing trial-and-error learning processes (Cavanagh et al., 2013; Guitart-Masip et al., 2012; Swart et al., 2018; Swart et al., 2017). More specifically, the Pavlovian bias could be elicited when a "no-go" instrumental response was required in the potential rewarding state or a "go" instrumental response in the potential losing state. It was found that medial-frontal theta oscillations are stronger when successfully overcoming the Pavlovian bias that conflicts with the instrumental behavior, indicating successful top-down control over Pavlovian bias during the instrumental learning process (Cavanagh et al., 2013; Swart et al., 2018).

Following our baseline report at age 18 (Chen et al., 2021d), we endeavored to test whether the behavioral performance along with the neural responses during the PIT task can predict the drinking trajectories of our sample during a six-year follow-up. Given that young

adulthood is a stage when drinking behavior escalates (Chen et al., 2004; Muthen & Muthen, 2000), increased alcohol consumption or binge drinking behavior during this stage may predispose the development of AUD in later stages of life. If the PIT effects were to be associated with increased alcohol use during young adulthood, it could potentially reflect a mechanism predisposing to AUD. To examine whether interference PIT effects can predict the drinking trajectories of our sample over 6 years (ages 18 to 24), we employed latent growth curve modeling. In addition to the PIT assessment at age 18, we included PIT data from one additional assessment that was assessed three years after study inclusion at age 18, i.e., at age 21.

We have previously reported an association between goal-directed and habitual control with risky drinking trajectories from ages 18 to 21 in this sample (Chen et al., 2021c). Consistent with the drinking trajectories modelled in this previous report, here, the first drinking trajectory of interest is an AUDIT-C trajectory (sum of the first three items of the Alcohol Use Disorders Identification Test), which assesses the frequency of drinking, the quantity of drinking in a typical drinking occasion, and the frequency of binge drinking since the last assessment. The second trajectory of interest is a binge drinking score trajectory that assesses the grams of ethanol intake during a typical drinking occasion. According to the World Health Organization (Stockwell et al., 2000), 60 g of ethanol or five standard drinks per drinking occasion can be considered the binge drinking threshold. However, this binary classification reduces dimensionality in the analysis, so the inclusion of a binge drinking score trajectory offers a continuous approach in assessing this behavior.

We assumed that a more substantial interference effect on the behavioral level and stronger neural response in the VS are associated with riskier drinking trajectories (i.e., positively associated with the slopes of the drinking trajectories); and that a weaker neural response in the lPFC and dmPFC would predict riskier drinking trajectories during the six-year follow-up period (i.e., showing negative associations with the slopes of the drinking trajectories).

## 4.3 Materials and Methods

### 4.3.1 Participants & General Procedure

The participants were recruited from the local registration offices in Berlin and Dresden (more details in Chen et al., [2021d]). At the beginning of the study, we included 201 males who are

right-handed and eligible for MRI, with neither history of nor current mental disorders, and with no substance dependence except for nicotine. The participants needed to have at least two drinking occasions during the last three months. Only males were recruited due to the predominance of male patients with AUD and dysfunctional alcohol consumption compared to female patients (Pabst & Kraus, 2008).

Participants performed the experimental procedure with two on-site appointments at baseline (age 18; *N* = 201) and the assessment three years later (age 21; *N* = 132). During the first appointment, participants completed the Munich-Composite International Diagnostic Interview (M-CIDI) (Jacobi et al., 2013; Wittchen & Pfister, 1997) based on the German version of the DSM-IV (Saß et al., 2003) and filled in other questionnaires that measure drinking-related behavior (descriptive statistics of the questionnaires of interest are displayed in C.5 in the Appendix); cognitive ability assessments including processing speed, working memory and crystalized intelligence were also performed (details in Chen et al., [2021d]). During the second appointment, participants performed the PIT task that consisted of four phases. The Pavlovian and PIT phases were done in the scanner, while the instrumental and forced-choice phases were conducted outside the scanner. The imaging data were acquired using a Siemens 3-Telsa MRI scanner (Magnetom Trio, Siemens, Erlangen, Germany). The details of the sequences and the preprocessing procedures are described in C.1 of the Appendix. After quality control (more information in Chen et al., [2021d]), 191 behavioral and 139 neural datasets were included for the baseline analysis.

Drinking behaviors were assessed over a six-year period from ages 18 to 24. In addition to the two on-site assessments, the participants were asked to fill in the AUDIT questionnaire online at 6-month intervals. Unfortunately, the AUDIT questionnaire was not available for the baseline assessment but only started six months after the baseline (at age 18.5). Besides the two on-site M-CIDI interviews at ages 18 and 21, M-CIDI telephone interviews were done every year when there were no on-site assessments. Regarding the main drinking behavior assessments that we analyzed for the current study, there were twelve AUDIT assessments (from ages 18.5 to 24; every 6 months) and seven M-CIDI interviews (ages 18-24; every year), which comprise of two on-site and five telephone interviews. In addition, participants needed to fill in other online questionnaires every year; more details about these assessments are

mentioned in the corresponding analyses, and the descriptive statistics are displayed in C.5 in the Appendix.

### 4.3.2 Alcohol Drinking Assessment

Consistent with our previous report on the three-year drinking trajectories (Chen et al., 2021c), we primarily focused on the AUDIT consumption score (AUDIT-C) and the gram/occasion variable from the M-CIDI interview. The AUDIT-C score was used to describe the alcohol consumption trajectory, given that it has been suggested to be sensitive to risky drinking and can be even more effective than the 10-item AUDIT total score (Dawson, 2011; Kuitunen-Paul et al., 2018). The gram/occasion variable from the M-CIDI interview assesses how many grams of alcohol the participants consume on a typical drinking occasion during the last year. As previously mentioned, this variable was used to measure the binge drinking behavior in a continuous way, as participants who continually consume more alcohol on a typical drinking occasion are more likely to be binge drinkers. Using a continuous variable instead of a binary categorization as binge and non-binge drinkers, we preserve more information in the variable, which also aligns with the DSM-V (Hasin et al., 2013) suggestions to characterize alcohol addiction with a more dimensional approach.

### 4.3.3 PIT Paradigm

The PIT paradigm is shown in Figure 15. This task has been described in more detail in the previous studies of our group (Garbusow et al., 2016; Garbusow et al., 2014).

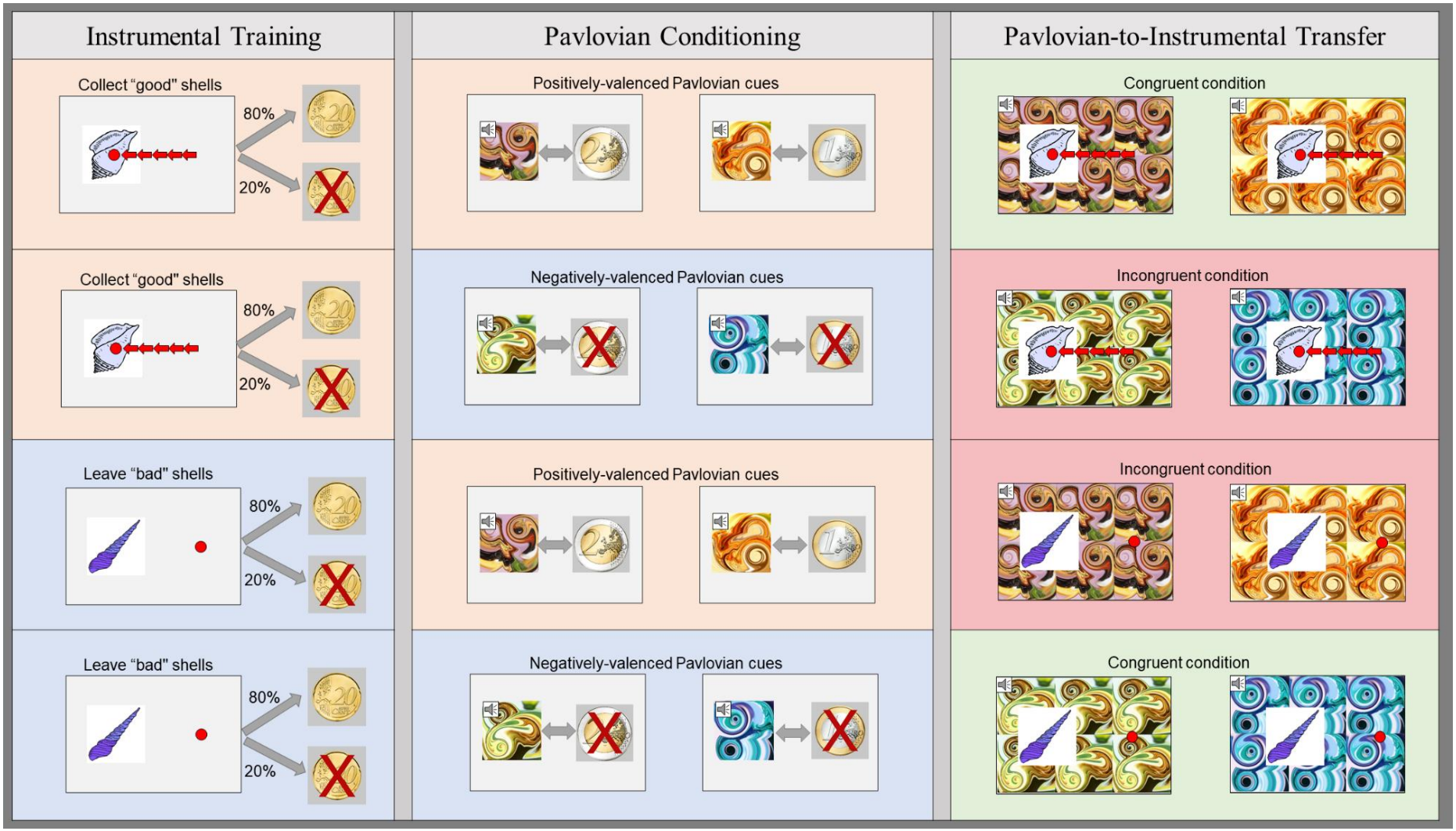
Instrumental Training
Pavlovian Conditioning
Pavlovian-to-Instrumental Transfer
Collect "good" shells
80%
20%
Positively-valenced Pavlovian cues
Congruent condition
Collect "good" shells
80%
20%
Negatively-valenced Pavlovian cues
Incongruent condition
Leave "bad" shells
80%
20%
Positively-valenced Pavlovian cues
Incongruent condition
Leave "bad" shells
80%
20%
Negatively-valenced Pavlovian cues
Congruent condition

**Figure 15: Pavlovian-to-instrumental transfer paradigm. Instrumental training**: Participants learned to collect good shells (press the button five or more times to move the dot toward the shell; colored in orange) and leave the bad shells (nothing needed to be done; colored in blue). A correct response yielded a €0.20 cent reward with the probability of 80% or a €0.20 cent monetary loss with a probability of 20%. After 60 trials, the instrumental training ended if the participants achieved the learning criterion (80% correct choices over 16 trials) or when a total number of 120 trials were reached. **Pavlovian conditioning:** Participants learned the association between five compound audiovisual stimuli (fractal images paired with pure tones) and the positive (€1, €2; colored in orange), neutral (€0) and negative (€-1, €-2; colored in blue) outcomes. The neutral condition is not shown in the figure since this condition cannot be categorized as "congruent" or "incongruent". The Pavlovian conditioning phase consisted of 80 trials, with each fractal appearing 16 trials. **Pavlovian-to-Instrumental Transfer phase:** Participants performed the instrumental task again with the fractal images tiled in the background; the pure tones were also played simultaneously. This phase was done in the MRI scanner and under nominal extinction to prevent further learning. Based on whether the Pavlovian background values were concordant with the instrumental stimulus or not, the experimental trials could be categorized into congruent (positively-valenced Pavlovian cues with “good” shells or negatively-valenced Pavlovian cues with “bad” shells; colored in green) and incongruent trials (positively-valenced Pavlovian cues with “bad” shells or negatively-valenced Pavlovian cues with “good” shells; colored in red). Each pairing of instrumental shell and Pavlovian cue appeared nine times during the transfer phase, resulting in 90 trials (9 trials × good/bad shells × five Pavlovian stimuli) in total. Among these trials, 36 trials belonged to the congruent, and 36 trials belonged to the incongruent conditions. Additionally, there were 72 trials during the transfer phase with water or alcohol pictures presented in the background. However, given that we have previously reported that the valence of water and alcohol backgrounds was perceived similarly to the neutral Pavlovian cue (Chen et al., 2021d), the alcohol/water trials along with the neutral trials were all excluded from the analyses.

### 4.3.4 Group-Level PIT Data Analysis

#### *4.3.4.1 Behavioral PIT Effect*

Eight subjects were excluded from the dataset at age 21 due to data loss caused by technical problems, leaving 124 complete datasets. Among these subjects, we excluded seven participants who did not have valid baseline data; therefore, 117 subjects who had valid PIT behavioral data at both ages 18 and 21 were included in the behavioral analyses. Consistent with the baseline

paper (Chen et al., 2021d), we calculated the difference in error rate between the incongruent condition and the congruent condition (∆ER) during the PIT phase at both ages 18 and 21. During the PIT phase, participants were instructed to perform the instrumental task according to what they had learned during the instrumental phase. Therefore, in the presence of the previously learned Pavlovian cues, the ∆ER variable reflected the extent to which individuals were susceptible to the influence of Pavlovian cues. Higher ∆ER values reflected more difficulty or inability to deal with the Pavlovian interference.

After characterizing the ∆ER as the behavioral PIT effect at both ages 18 and 21, we first calculated the Pearson's correlation between them. This tested how strongly the behavioral PIT effects from the two time points were associated and can also indicate the test-retest reliability. A paired sample *t* test was done to investigate whether there were significant changes in the behavioral PIT effect over the three years on the group level. The ∆ER from ages 18 and 21 were then used as the two PIT behavioral predictors to predict the individual drinking trajectories.

*4.3.4.2 Neural PIT Effect*

Regarding the neural data at age 21, we excluded four participants who had missing imaging data and four more participants who had either more than 3 mm volume-to-volume movement or more than 3° rotation. After further excluding those participants who did not have valid baseline neural data, 79 subjects with valid neural data from ages 18 and 21 remained for the fMRI analyses. The data analyses were performed with SPM12 (Wellcome Trust Centre for Neuroimaging, London, UK). The first- and second-level models were constructed in the same way as our baseline paper (Chen et al., 2021d). More specifically, mirroring our behavioral analysis, the incongruent versus congruent contrast was defined individually as the first-level model. This contrast was then entered into the second-level analysis as a one-sample *t* test. The individual behavioral PIT interference effect (∆ER) was included as the covariate on the second level; the site information (whether the experiment was performed in Berlin or Dresden) was additionally included as a covariate of no interest to control for the potential site differences (described with more details in the Appendix C.3).

At baseline, we have shown that the neural responses in the VS, lPFC, and dmPFC in the incongruent versus congruent contrast were positively associated with the behavioral PIT effect (Chen et al., 2021d). Following this, we now first investigated neural correlates of the behavioral PIT interference effect at the whole-brain level with an uncorrected threshold of $p < .001$ and a cluster size $k \geq 50$. We then extracted the parameter estimates during the incongruent condition within the same sets of regions of interest (ROI) to obtain neural PIT predictors from both ages (i.e., 18 and 21) for the latter drinking trajectory analysis. We chose to extract the neural responses in the ROIs during the incongruent trials as the main neural predictors to predict the drinking trajectories, since these neural correlates of the PIT interference effect had been found to be associated with risk status at age 18 in our baseline report (Chen et al., 2021d). The three ROIs, including the VS, lPFC, and dmPFC, were defined based on previous meta-analyses (details in C.4 in the Appendix). We did not extract the parameter estimates from the amygdala since no association with the interference PIT effect was found in the baseline analyses. After extracting the parameter estimates from the three ROIs, we again calculated the Pearson's correlation coefficients between neural responses at ages 18 and 21 to check whether the neural responses within the three ROIs were reliable. Additionally, we performed paired sample *t* tests to check whether there were significant changes in the neural responses during the incongruent condition across the three years on the group level.

### 4.3.5 Group-Level Drinking Behavior Analysis

To gain an impression of the drinking behavior on the group level, we first plotted the histograms of both variables (Figure 16). Regarding the AUDIT-C development on the group level (Figure 16B), there seemed to be a minor decrease over time; we thus regressed this variable against time (as a categorical variable) to test whether this decrease was significant. According to Figure 16D, on average, the gram/occasion variable first decreased and then increased. Therefore, we regressed this variable against both a linear term and a quadratic term (squared time; $time^2$) to check whether the increase and decrease were significant on the group level.

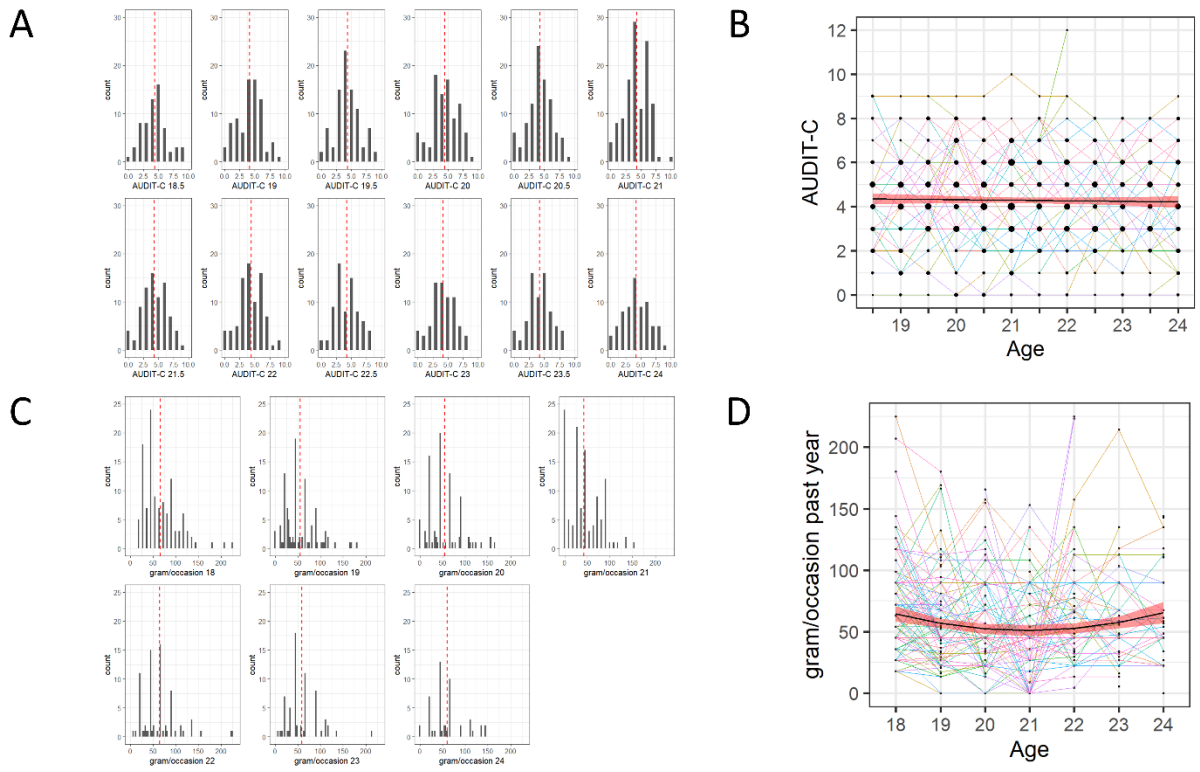


**Figure 16. Histograms and individual trajectories**. A and C: Histograms from all available measurements for the Alcohol Use Disorders Identification Test consumption score (AUDIT-C) and gram/occasion variables. The group means are indicated by the red dashed lines. B and D: Individual trajectories, shown in different colors, are plotted against age. Group means are shown with the bold, solid lines, and the red areas around the group mean lines indicate standard error.

### 4.3.6 Individual Drinking Trajectory Analysis

The latent growth curve modeling (LGCM) approach offers a multi-level framework that investigates both intra- and inter-individual changes in longitudinal studies. On the first level (intra-individual level), individual intercepts and slopes can be used to characterize the intra-individual developmental trajectories when linearity is assumed. A quadratic slope could also be added to the first-level model if a quadratic developmental pattern is assumed. On the second level, one can include different predictors in the model to investigate the association between these predictors and individual intercepts, linear, and quadratic slopes. We implemented the LGCM analyses with the lavaan package in R Studio (Rosseel, 2012). With the lavaan package, the

missing data could be handled via the full information maximum likelihood method, where likelihood functions are estimated for the individuals according to the available information. Importantly, when assuming the missing data to be random, this method is suggested to be unbiased (Arbuckle et al., 1996).

To investigate how the interference PIT effects were associated with the development of risky drinking behavior, we created two drinking trajectories with the variables of interest: AUDIT-C and gram/occasion. Before including any predictors, we first compared models with a linear slope, a quadratic slope, and linear + quadratic slopes to decide which best described the intra-individual drinking trajectories. We compared the linear and quadratic slope model with the linear + quadratic slope model with chi-squared tests, given that they are nested models. When the linear and quadratic slope models showed a better fit than the linear + quadratic slope model, we based our final decision on each model's Bayesian Information Criterion (BIC). The model with the lowest BIC was determined to be the superior model. Additionally, we computed the correlation between the intercepts and slopes from the two unconditional drinking trajectory models to check whether the two drinking trajectories developed differently over time.

In the next step, we included the PIT predictors into the best-fitting trajectory model. We built separate models with either behavioral or neural PIT effects as the predictors for the two drinking trajectories (four models in total). In order to preserve more behavioral data sets, we did not include all predictors (behavioral and neural) into one model, as done in Chen et al. (2021c). Otherwise, it would have meant that only 79 subjects who had complete behavioral and neural data could be included in the behavioral analysis. The behavioral model for the AUDIT-C trajectory is displayed in Figure S9A in the Appendix. The figure shows that all the behavioral PIT paths at age 18 to the intercept and slopes were freely estimated. We only included the paths from the behavioral PIT effect at 21 to the slopes but not to the intercepts because this PIT assessment occurred later than the baseline drinking behavior. The covariance structures between ages 18 and 21 were also freely estimated. The same model structure was specified for the binge drinking score model.

The neural PIT models were constructed following the same line of reasoning as the behavioral PIT model (see Figure S9B for the AUDIT-C model). Specifically, we included the VS, lPFC, and dmPFC neural responses during the incongruent trials at ages 18 and 21 as six neural predictors. Compared to our baseline report, where we used the data-driven activated clusters, we used the neural responses from the ROIs to maintain consistency between the two assessments at ages 18 and 21. The paths from the three baseline neural predictors to the intercept and slopes of the drinking trajectories were freely estimated. For the three neural predictors at age 21, the paths were again only directed from the neural predictors to the slopes. Additionally, covariance structures were estimated between all pairs of neural responses at the same time point and between the neural responses within the same ROI across the two assessments.

### 4.3.7 Exploratory Analyses

To better understand the association between the behavioral PIT effect and the AUDIT-C trajectories, we performed cluster analyses to identify distinctive developmental patterns. The behavioral PIT effect at the two assessments, as well as the change in the behavioral PIT effect from age 18 to 21 were also compared between the clusters. We additionally explored whether other questionnaires of interest (descriptive statistics in the Appendix C.8) could characterize the cluster profiles through logistic regression. We described these exploratory analyses together with the motivation behind it in details in the Appendix C.6.

Further, to gain more insights into the difference between AUDIT-C and gram/occasion variables, we calculated the correlation coefficients between AUDIT-C, gram/occasion, and the obsessive compulsive drinking scale (OCDS) total score (Anton et al., 1995; Mann & Ackermann, 2000), as well as the alcohol dependence scale (ADS) (Skinner & Allen, 1982) sum score whenever they were assessed at the same time point. The motivation of this analysis is explained in more details in Appendix C.7.

## 4.4 Results

### 4.4.1 Behavioral PIT Effect on the Group Level

Recently, we reported the behavioral PIT effect in 191 participants (Chen et al., 2021d). The current study only reports the behavioral PIT effects of the 117 participants who performed the PIT task at both ages 18 and 21. At age 18, the 117 participants showed an increase in ER by 15.1% in the incongruent condition compared to the congruent condition on average ($T$ = 5.58; df = 116; $p = 1.63\times10^{-7}$; Cohen's d = 0.52). At age 21, a similar pattern was found. The ER in the incongruent condition was increased by 17.0% compared to the congruent condition ($T$ = 5.72; df = 116; $p = 8.41\times10^{-8}$; Cohen's d = 0.53). The correlation between the behavioral PIT effects at ages 18 and 21 was significant (r(115) = 0.29, $p$ = .002). There was no significant change in the behavioral PIT effect across the two assessments as indicated by a paired-sample $t$ test ($T$ = -0.56, df = 116, $p$ = .578; Cohen's d = 0.07).

### 4.4.2 Neural PIT Effect on the Group Level

In the initial analysis of 139 participants, we found a neural interference PIT effect in the VS, lPFC, and dmPFC (Chen et al., 2021d). Now, we analyzed a subsample of 79 participants who had valid neural PIT data at both time points. At age 18, we found that the neural PIT effect of this subsample was comparable to the neural activation pattern previously reported with the 139 subjects. As shown in Figure 17A, the neural PIT effect was found in the caudate (extended to the ventral striatum; $k$ = 74, $T$ = 4.04, peak MNI coordinate: 12/16/2), lPFC, and dmPFC (within the same cluster; $k$ = 8,109, $T$ = 5.53, peak MNI coordinate: 8/18/48) with an uncorrected whole-brain threshold of $p < .001$ and a cluster size of $k \geq 50$. Using the same threshold, the neural PIT effect at age 21 was again found in the caudate (also extended to the ventral striatum; $k$ = 465, $T$ = 4.64, peak MNI coordinate: 14/14/8) and lPFC ($k$ = 73, $T$ = 3.73, peak MNI coordinate: 38/52/10). The analysis also revealed a cluster in the anterior cingulate cortex ($k$ = 61, $T$ = 3.62, peak MNI coordinate: 6/44/18) (Figure 17B). Since the neural activation was not found in the dmPFC, one of our ROIs, we used a lower threshold of $p < .005$ and confirmed that the anterior cingulate cortex extended to the dmPFC mask we applied in the initial report.

The correlation between the neural responses in the ROIs during the incongruent trials at ages 18 and 21 was moderate for the VS ($r(77) = 0.43$, $p = 8.03\times10^{-5}$) and weak for lPFC ($r(77) = 0.29$, $p = .009$) and dmPFC ($r(77) = 0.33$, $p = .003$). According to the paired-sample $t$ tests ($p > 0.53$), there were no significant changes between the neural responses in the incongruent condition between ages 18 and 21.

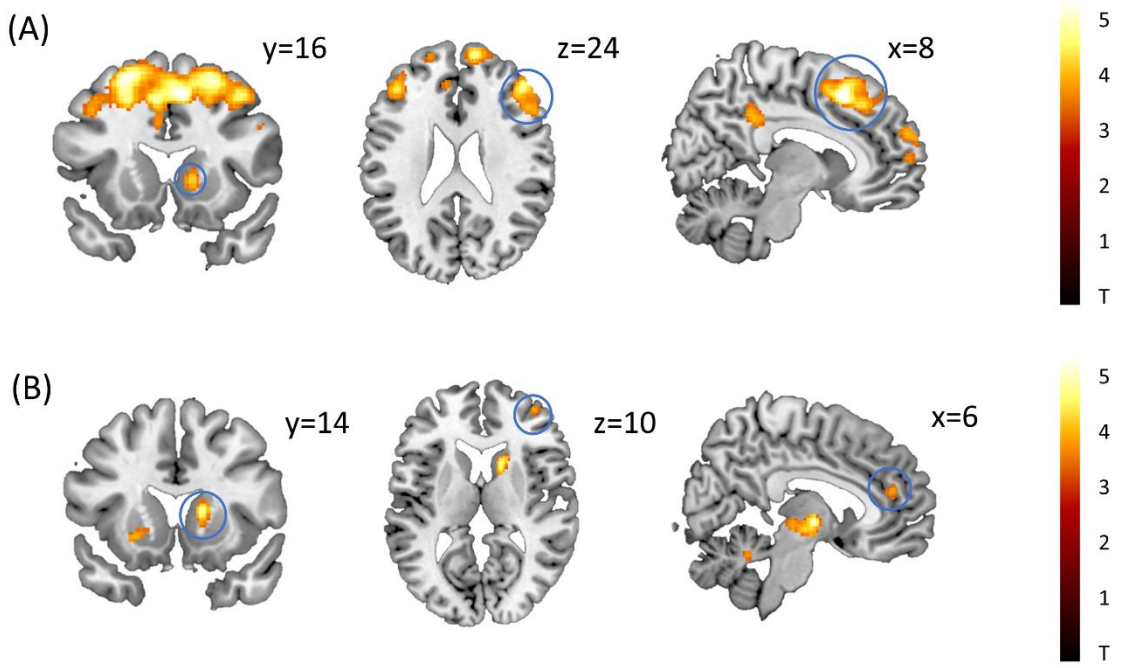


**Figure 17: Neural Pavlovian-to-instrumental transfer (PIT) results**. (A) Neural interference PIT effect (neural responses during interference correlated with the behavioral PIT effect) at baseline was found in the caudate (extended to the ventral striatum; $k = 74$, $T = 4.04$, peak MNI coordinate: 12/16/2), the lateral prefrontal and the dorsomedial prefrontal cortices (within the same cluster; $k = 8{,}109$, $T = 5.53$, peak MNI coordinate: 8/18/48); displayed with the threshold of $p < .001$, cluster size $k \geq 50$ ($N = 79$). (B) Neural interference PIT effect at age 21 was found in the caudate (also extended to the ventral striatum; $k = 465$, $T = 4.64$, peak MNI coordinate: 14/14/8), the lateral prefrontal cortex ($k = 73$, $T = 3.73$, peak MNI coordinate: 38/52/10), as well as anterior cingulate cortex ($k = 61$, $T = 3.62$, peak MNI coordinate: 6/44/18); displayed with the same threshold.

### 4.4.3 Drinking Behavior on the Group Level

On the group level, there was no significant change in the AUDIT-C over the six year period ($\beta$ = -0.01, $p$ = .54), with the mean AUDIT-C score ranging between 4.03 and 4.44 across the six years. Concerning the gram/occasion variable, the statistical analysis confirmed that this variable first decreased and increased as time passed ($\beta$ = -12.27; $p$ < .001 for the linear term; $\beta$ = 1.56; $p$ < .001 for the quadratic term). On average, the mean alcohol intake per drinking occasion decreased from 66 to 43 g from ages 18 to 21, and then increased to 60 g at age 24. The descriptive statistics of the two variables of interest, along with other drinking-related variables are shown in C.2 in the Appendix. Importantly, as shown in the trajectory plots of Figure 16, individuals showed different patterns in their drinking trajectories regardless of whether or not there was a change on the group level.

### 4.4.4 Individual Drinking Trajectory Model Comparisons

The next step was to select the best fitting model for the six-year trajectories at the individual level. Regarding the AUDIT-C trajectory, we found that adding a quadratic term improved the model fit when compared with the just the linear slope model ($\Delta\chi2$ (3, 117) = 18.95, $p$ < .001), as well as compared to an only quadratic term model ($\Delta\chi2$ (3, 117) = 14.54, $p$ = .002). Therefore, the linear + quadratic model was chosen for the AUDIT-C trajectory. Essentially, when formalizing trajectories with a linear and a quadratic term, we can capture more types of developmental courses compared to a linear function only, which best fits constant decreases, no changes, or constant increases over time. For example, when the linear term is negative, and the quadratic term is positive, the drinking behavior decreases initially; later on, the positive quadratic term can contribute to an increase of the drinking behavior after a turning point. To demonstrate how the different combinations of intercept, linear and quadratic slopes can lead to different trajectories, we plotted six examples of AUDIT-C trajectories (C.10 in the Appendix).

The model comparison for the gram/occasion trajectory showed that the linear slope model was slightly better than the linear + quadratic trajectory model ($\Delta\chi2$ (3, 117) = 7.61, $p$ = .055). Still, the quadratic slope model showed a worse fit than the linear + quadratic slope model ($\Delta\chi2$ (3, 117) = 9.48, $p$ = .024). Considering the linear slope model had the lowest BIC (6164.8) compared to the

quadratic (6166.7) and the linear + quadratic slope (6171.5) models, we chose the linear slope model for the gram/occasion trajectory.

Following this, we extracted the individual intercepts and slopes from the two drinking trajectory models and investigated their associations with each other. There were high correlations between the intercepts of the two variables ($r(115) = 0.61$, $p < .001$). However, the linear slope of the gram/occasion showed a weak association with the linear slope of the AUDIT-C ($r(115) = 0.10$, $p = .282$) and almost no association with the AUDIT-C quadratic slope ($r(115) = -0.03$, $p = .762$), indicating that they developed differently over the six years.

### 4.4.5 Individual Behavioral Models

We then tested whether the PIT behavioral interference effect at ages 18 and 21 was associated with the linear and quadratic slopes of the AUDIT-C trajectory. The AUDIT-C model showed a good model fit ($\chi 2 = 114.39$, df = 79, $p = .006$, CFI = 0.972, RMSEA = 0.062, SRMR = 0.050). Among all the regressions of interest, we found that the behavioral PIT effect at age 21 was negatively associated with the linear slope ($\beta = -0.351$; $p = .005$) but positively associated with the quadratic slope ($\beta = 0.031$, $p = .009$). The results are shown in Figure 18, and the path estimates of other associations are displayed in Table 5.

The negative association between the behavioral PIT effect at age 21 and the linear slope suggests that a stronger behavioral PIT effect was initially associated with a stronger decrease. Since the quadratic term drives the upward trend after the turning point, the positive association with the quadratic slope indicates that participants with a stronger behavioral PIT effect showed a more substantial increase after the turning point. To better visualize the association between the behavioral PIT effect and both the linear and quadratic trajectories, we additionally plotted the standardized estimates to show how one or two standard deviations from the group mean of PIT behavioral effect could be projected to the AUDIT-C trajectory development (Figure 20A). As seen in Figure 20A, since the turning point was around age 21, a stronger behavioral PIT effect at age 21 was associated with a steeper decrease from ages 18 to 21 but a more pronounced increase in drinking behaviors after age 21.

The gram/occasion model fit was acceptable ($\chi 2 = 45.39$, df = 29, $p = .027$, CFI = 0.886, RMSEA = 0.070, SRMR = 0.100), and we did not find any significant associations between the PIT behavioral effects with the gram/occasion trajectory (results displayed in Table 5).

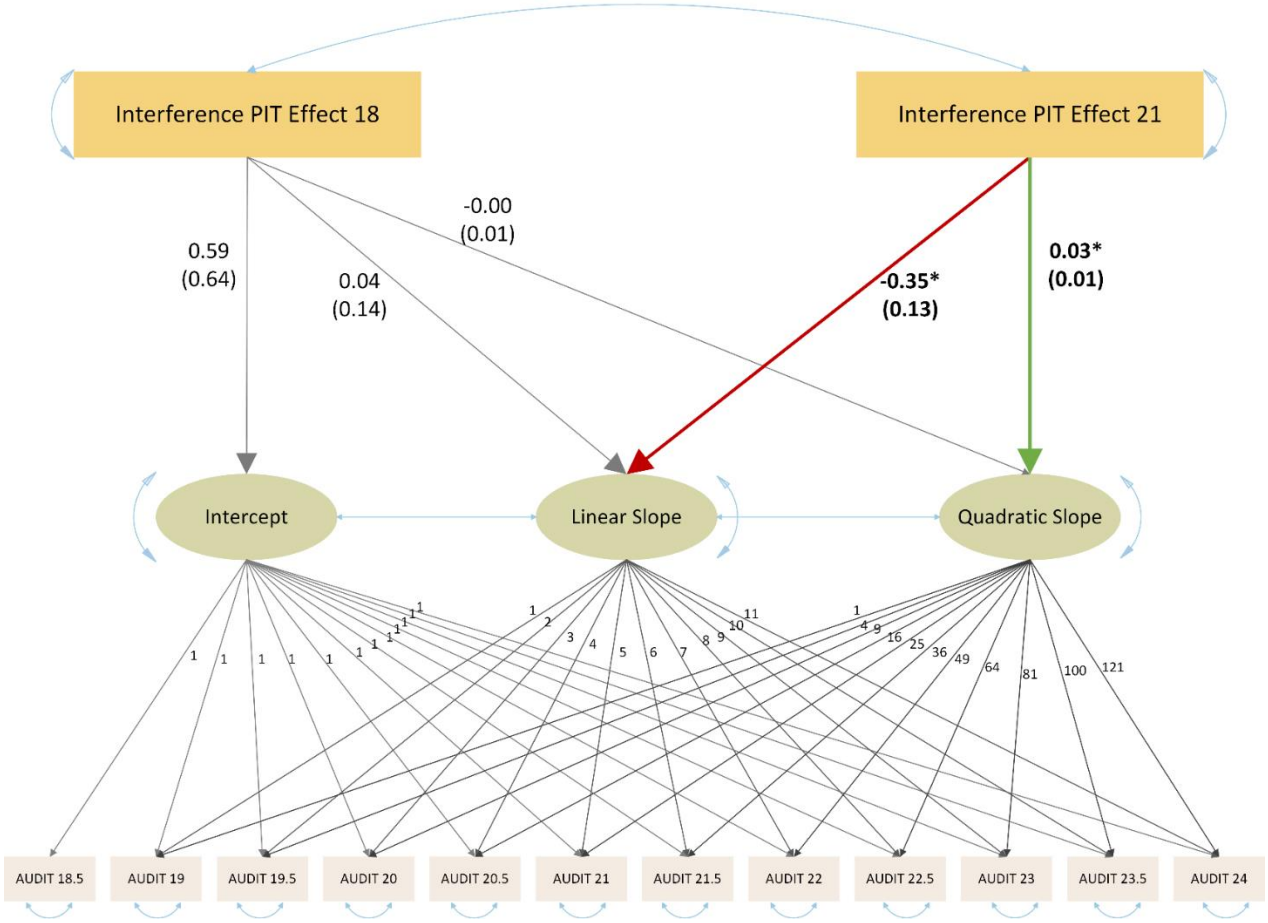


**Figure 18: Behavioral latent growth curve model for the Alcohol Use Disorders Identification Test consumption score (AUDIT-C) trajectory**. The observed variables are displayed within rectangles; the blue double-headed arrows specify the estimated variances. Three latent variables (intercept, linear, and quadratic slopes) were created for the AUDIT-C model, with the fixed loadings shown along the paths. The path estimates are also displayed in the figure. It was found that the behavioral PIT effect at age 21 was negatively associated with the linear slope (red path) but positively associated with the quadratic slope (green path).

### 4.4.6 Individual Neural Models

The AUDIT-C neural model showed a good model fit ($\chi 2 = 223.39$, df = 124, $p < .001$, CFI = 0.912, RMSEA = 0.101, SRMR = 0.102). As shown in Figure 19, the only significant association we found at age 18 was a positive association between the neural response in the VS and the intercept of the AUDIT-C trajectory ($\beta = 0.186$, $p = .010$). At age 21 dmPFC responses were positively associated with the quadratic slope ($\beta = 0.003$, $p = .043$). Further inspection of this effect showed that this association between the dmPFC neural responses at age 21 and the AUDIT-C trajectory shared a very similar, albeit statistically weaker, pattern compared to the behavioral PIT results (see Figure 20B).

The gram/occasion model showed an acceptable model fit ($\chi 2 = 72.03$, df = 57, $p = .087$, CFI = 0.953, RMSEA = 0.058, SRMR = 0.124). The VS response during incongruent condition at age 18 was positively associated with the intercept of this trajectory ($\beta = 3.120$, $p = .004$) but negatively associated with the linear slope ($\beta = -0.535$, $p = .035$).

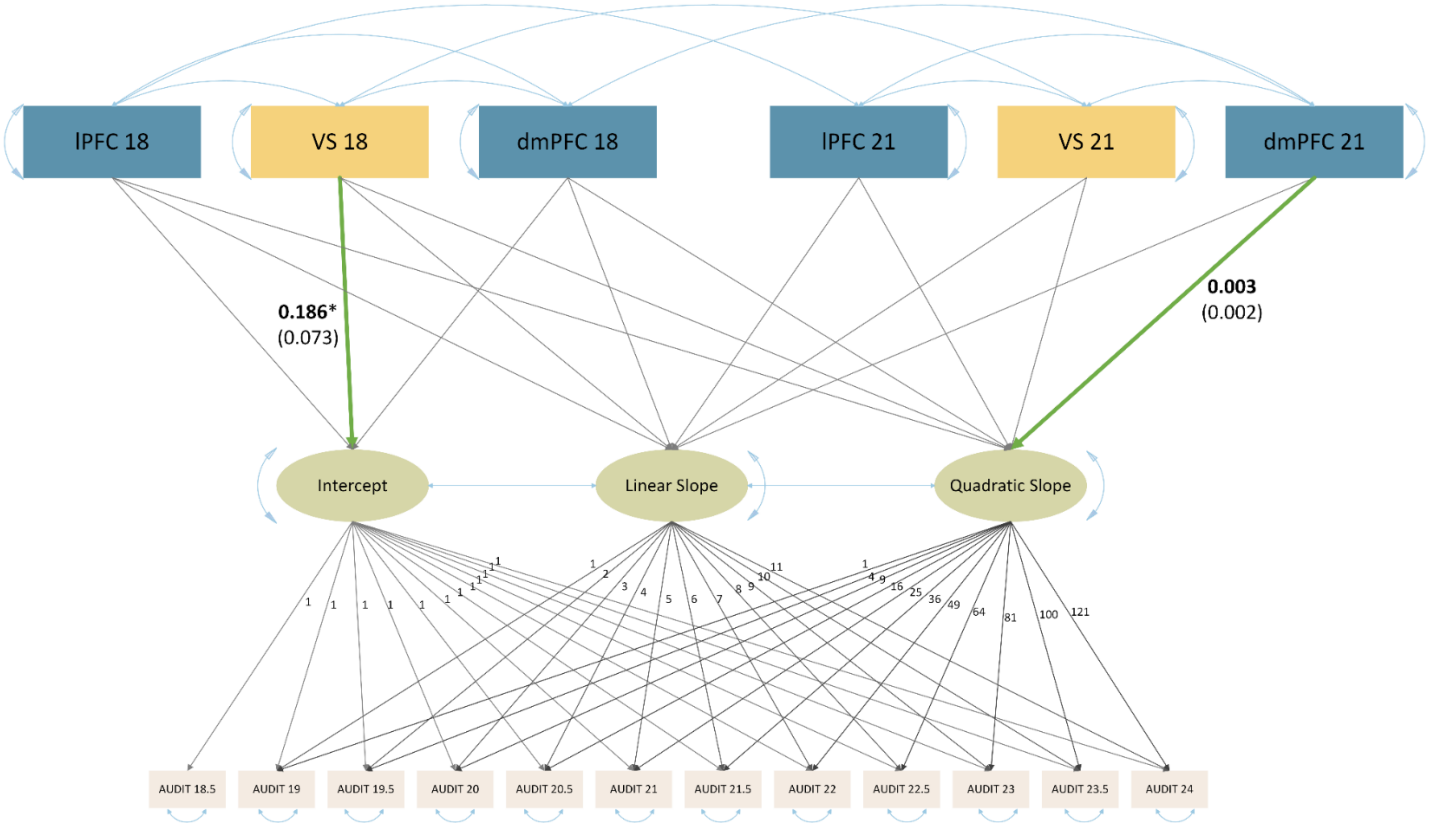


**Figure 19: Neural latent growth curve model for Alcohol Use Disorders Identification Test consumption score (AUDIT-C) trajectory.** The observed variables are shown in rectangles. The ventral striatum (VS), lateral prefrontal cortex (lPFC), and dorsomedial prefrontal cortex (dmPFC) responses during the incongruent trials at ages 18 and 21 were used as predictors. The loadings from the intercept, linear, and quadratic slopes to the AUDIT-C were fixed. Other regressions and covariance as indicated by the blue arrows were freely estimated. The bold green path (left) showed that there was a positive association between the VS response in the incongruent trials at age 18 and the intercept of the AUDIT-C trajectory. Additionally, the dmPFC responses during incongruent condition were positively associated with the quadratic slope (right green path). For the readability of the graph, we only showed the significant paths; the estimates of the paths that did not show significant effects are displayed in Table 5.

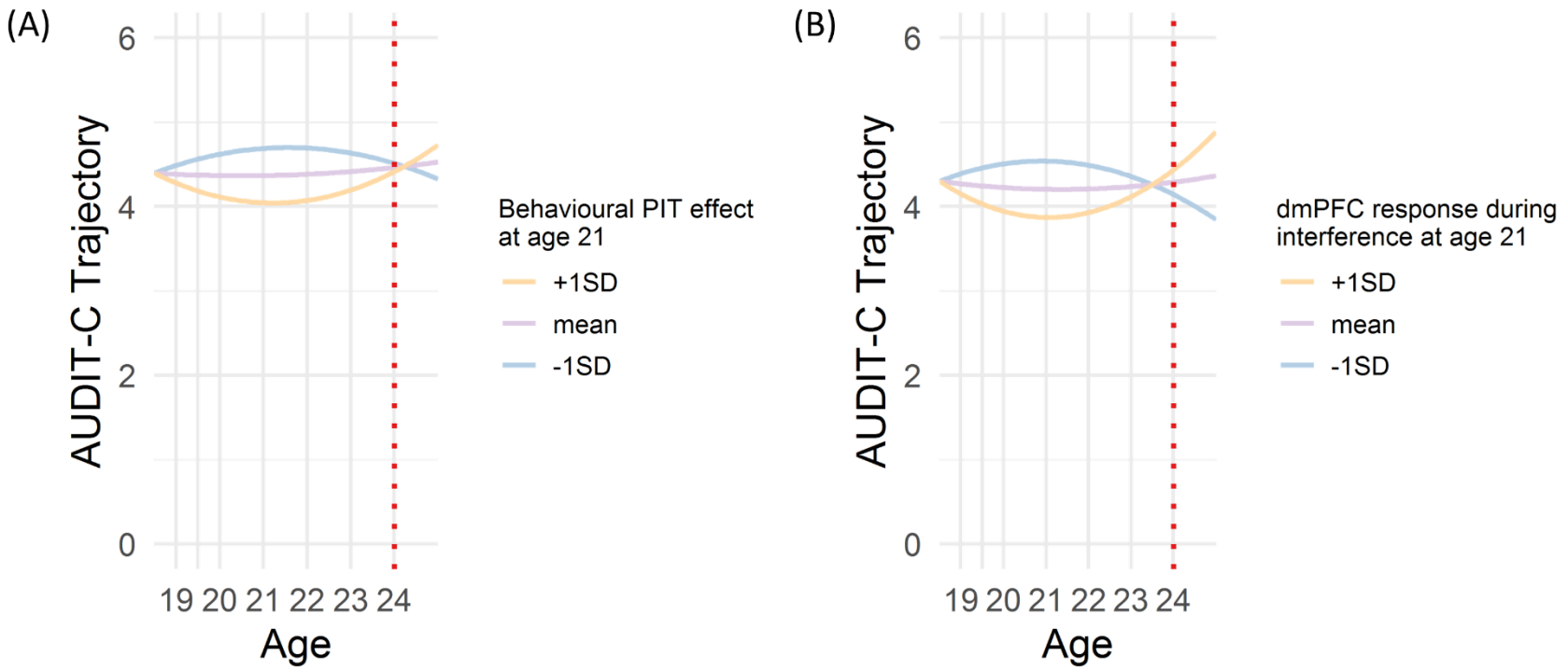


**Figure 20: Illustration of the association between the behavioral Pavlovian-to-instrumental transfer (PIT) effect and the dorsomedial prefrontal cortex (dmPFC) neural responses during incongruent trials at age 21 and the Alcohol Use Disorders Identification Test consumption score (AUDIT-C) quadratic trajectory**. The three lines specify how the AUDIT-C trajectories develop when the PIT behavioral effect or dmPFC neural responses at age 21 are at the group mean as well as one standard deviation (SD) below or above the group mean. In order to plot this effect, we centered all the variables and re-estimated the behavioral and neural AUDIT-C models. In this way, the mean estimates of intercept, linear and quadratic slopes indicate the trajectories where the behavioral PIT effect and dmPFC neural responses were set at the group mean (AUDIT-C behavioral trajectory = intercept + linear slope × t + quadratic slope × $t^2$ = 4.397 - 0.016 × t + 0.002 × $t^2$; AUDIT-C neural trajectory = 4.300 - 0.034 × t + 0.003 × $t^2$). For the trajectories at one SD below or above the group mean, the linear and quadratic slopes were adjusted according to the change that is associated with one SD change in the behavioral PIT effect or dmPFC responses ($SD_{PIT}$ ×

[path estimate PIT → linear/quadratic slope]). Since neither the behavioral PIT effect nor the dmPFC neural responses at age 21 were assumed to be associated with the intercept, we used a fixed starting point according to the intercept estimate. Specifically, it is plotted with the following equation: AUDIT-C behavioral trajectory = intercept + (linear slope ± $SD_{PIT}$ × [path estimate behavioral PIT → linear slope]) × t + (quadratic slope ± $SD_{PIT}$ × [path estimate behavioral PIT → quadratic slope]) × $t^2$ = 4.397 + (- 0.016 ± 0.322 × (-0.355)) × t + (0.002 ± 0.322 × 0.031)) × $t^2$; AUDIT-C neural trajectory = 4.300 + (-0.034 ± 4.451 × (-0.030)) × t + (0.003 ± 4.451 × 0.003)) × $t^2$.

*Table 5: LGCM results*

| | | Path | Estimate | SE | Estimate | *Z* | *P* value | Explained Variances |
|---|---|---|---|---|---|---|---|---|
| | | | (unstandardized) | | (standardized) | | | |
| **AUDIT consumption score** | | | | | | | | |
| Behavioral | delta ER - 18 | delta ER → intercept | 0.588 | 0.638 | 0.096 | 0.922 | .357 | 2.13% |
| | | delta ER → linear slope | 0.044 | 0.142 | 0.047 | 0.308 | .758 | 0.94% |
| | | delta ER → quadratic slope | -0.004 | 0.013 | -0.042 | -0.282 | .778 | 0.18% |
| | delta ER - 21 | delta ER → linear slope | -0.351 | 0.126 | -0.412 | -2.780 | **.005** | **16.97%** |
| | | delta ER → quadratic slope | 0.031 | 0.012 | 0.392 | 2.602 | **.009** | **19.54%** |
| Neural | VS - 18 | VS → intercept | 0.186 | 0.073 | 0.420 | 2.560 | **.010** | **22.09%** |
| | | VS → linear slope | -0.018 | 0.019 | -0.240 | -0.958 | .338 | 5.76% |
| | | VS → quadratic slope | 0.001 | 0.002 | 0.092 | 0.361 | .718 | 2.02% |
| | lPFC - 18 | lPFC → intercept | -0.035 | 0.076 | -0.101 | -0.460 | .646 | 1.02% |
| | | lPFC → linear slope | -0.017 | 0.019 | -0.285 | -0.874 | .382 | 8.12% |
| | | lPFC → quadratic slope | 0.001 | 0.002 | 0.284 | 0.838 | .402 | 11.16% |
| | dmPFC - 18 | dmPFC → intercept | -0.069 | 0.071 | -0.214 | -0.975 | .329 | 4.58% |
| | | dmPFC → linear slope | 0.022 | 0.019 | 0.394 | 1.167 | .243 | 19.71% |
| | | dmPFC → quadratic slope | -0.001 | 0.002 | -0.233 | -0.671 | .502 | 5.43% |
| | VS - 21 | VS → linear slope | 0.019 | 0.019 | 0.219 | 1.036 | .300 | 7.24% |
| | | VS → quadratic slope | -0.002 | 0.002 | -0.209 | -0.910 | .363 | 4.37% |
| | lPFC - 21 | lPFC → linear slope | 0.008 | 0.018 | 0.115 | 0.482 | .630 | 2.72% |
| | | lPFC → quadratic slope | -0.001 | 0.002 | -0.193 | -0.744 | .457 | 3.72% |
| | dmPFC - 21 | dmPFC → linear slope | -0.031 | 0.017 | -0.430 | -1.857 | .063 | 18.49% |
| | | dmPFC → quadratic slope | 0.003 | 0.002 | 0.506 | 2.027 | **.043** | **30.91%** |

| Binge drinking score (gram alcohol / drinking occasion) past year | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Behavioral | delta ER - 18 | delta ER → intercept | 14.209 | 10.429 | 0.155 | 1.362 | .173 | 4.20% |
| | | delta ER → linear slope | -1.592 | 2.670 | -0.089 | -0.596 | .551 | 0.79% |
| | delta ER - 21 | delta ER → linear slope | 0.512 | 1.902 | 0.031 | 0.269 | .788 | 0.66% |
| Neural | VS - 18 | VS → intercept | 3.120 | 1.079 | 0.508 | 2.890 | **.004** | **31.14%** |
| | | VS → linear slope | -0.535 | 0.254 | -0.554 | -2.109 | **.035** | **30.69%** |
| | lPFC - 18 | lPFC → intercept | 0.088 | 1.138 | 0.018 | 0.077 | .938 | 0.46% |
| | | lPFC → linear slope | 0.036 | 0.282 | 0.009 | 0.129 | .898 | 0.35% |
| | dmFPC - 18 | dmPFC → intercept | -1.590 | 1.038 | -0.355 | -1.532 | .126 | 12.60% |
| | | dmPFC → linear slope | 0.269 | 0.253 | 0.381 | 1.061 | .289 | 18.58% |
| | VS - 21 | VS → linear slope | 0.417 | 0.237 | 0.371 | 1.759 | .079 | 17.72% |
| | lPFC - 21 | lPFC → linear slope | 0.052 | 0.210 | 0.056 | 0.249 | .803 | 1.12% |
| | dmFPC - 21 | dmPFC → linear slope | -0.037 | 0.191 | -0.041 | -0.196 | .845 | 0.17% |

### 4.4.7 Results of the AUDIT-C Clustering Analysis

As described in Appendix C.6, we conducted the clustering analysis based on the linear and quadratic slopes, using a fixed cluster number of two. The first cluster had a positive linear but negative quadratic slope, and vice versa for the second cluster. The mean trajectories of the two clusters (Figure 21A) reveal that the first cluster peaked around age 21 and decreased afterwards. In contrast, the second cluster first decreased and then developed prominently until or further beyond age 24. We thus labelled the two clusters as "early peaker" ($N$ = 59) and "late riser" ($N$ = 58) group, respectively.

When comparing the behavioral PIT effect between the two subgroups (Figure 21B and C), we found that the two groups did not show any differences in the behavioral PIT effect at age 18 ($T$ = -0.30, df = 114, $p$ = .765), but at age 21: The “late riser” group showed a 3-times higher interference PIT effect as compared to the “early peaker” group ($T$ = -3.27, df = 105, $p$ = .001). These results are in line with the LGCM analysis. Further, they suggested that the association between the behavioral effect and the linear, as well as quadratic slopes, were mainly driven by the “late riser” group. Moreover, as displayed in Figure 21D, the change of the PIT effect from age 18 to 21 was different between the two groups ($T$ = -2.58, df = 114, $p$ = .011): the late risers showed a significant increase in the PIT effect ($T$ = 2.14, df = 57, $p$ = .037), while the “early peakers” seemed to show a nominal decline, though this change was not significantly different from zero ($T$ = -1.48, df = 58, $p$ = .146).

Mirroring this pattern on the behavioral level, the “late riser” group, as compared with the “early peaker” group, showed stronger dmPFC responses during conflict at age 21 ($T$ = -2.43, df = 77, $p$ = .017), but neither a significant difference to dmPFC responses at age 18 ($T$ = -0.88, df = 75, $p$ = .380) nor different changes in dmPFC responses from ages 18 to 21 ($T$ = -0.85, df = 69, $p$ = .398). These effects are depicted in Figure 21E-G. Conversely, the two groups did not differ regarding their VS and lPFC responses during conflict at age 18 or 21; changes in neural responses from ages 18 to 21 within these two regions were not significant ($p$ > .099).

Through logistic regression, we explored whether other questionnaires of interest, in addition to the behavioral PIT effect at age 21, could explain why people belong to different subgroups

(described in Appendix C.6). The logistic regression showed that, in addition to the behavioral PIT effect at age 21, a stronger social motive to consume alcohol at age 21 ($\beta$ = 0.38, $p$ = .037) and higher socioeconomic status at age 18 were associated with a higher likelihood of being in the "late riser" group. Conversely, more physical neglect during childhood ($\beta$ = -0.79; $p$ = .027) and higher alexithymia score ($\beta$ = -0.13, $p$ = .045) were associated with the membership of "early peaker" group. The complete results of the logistic regression are displayed in Table S12 in the Appendix.

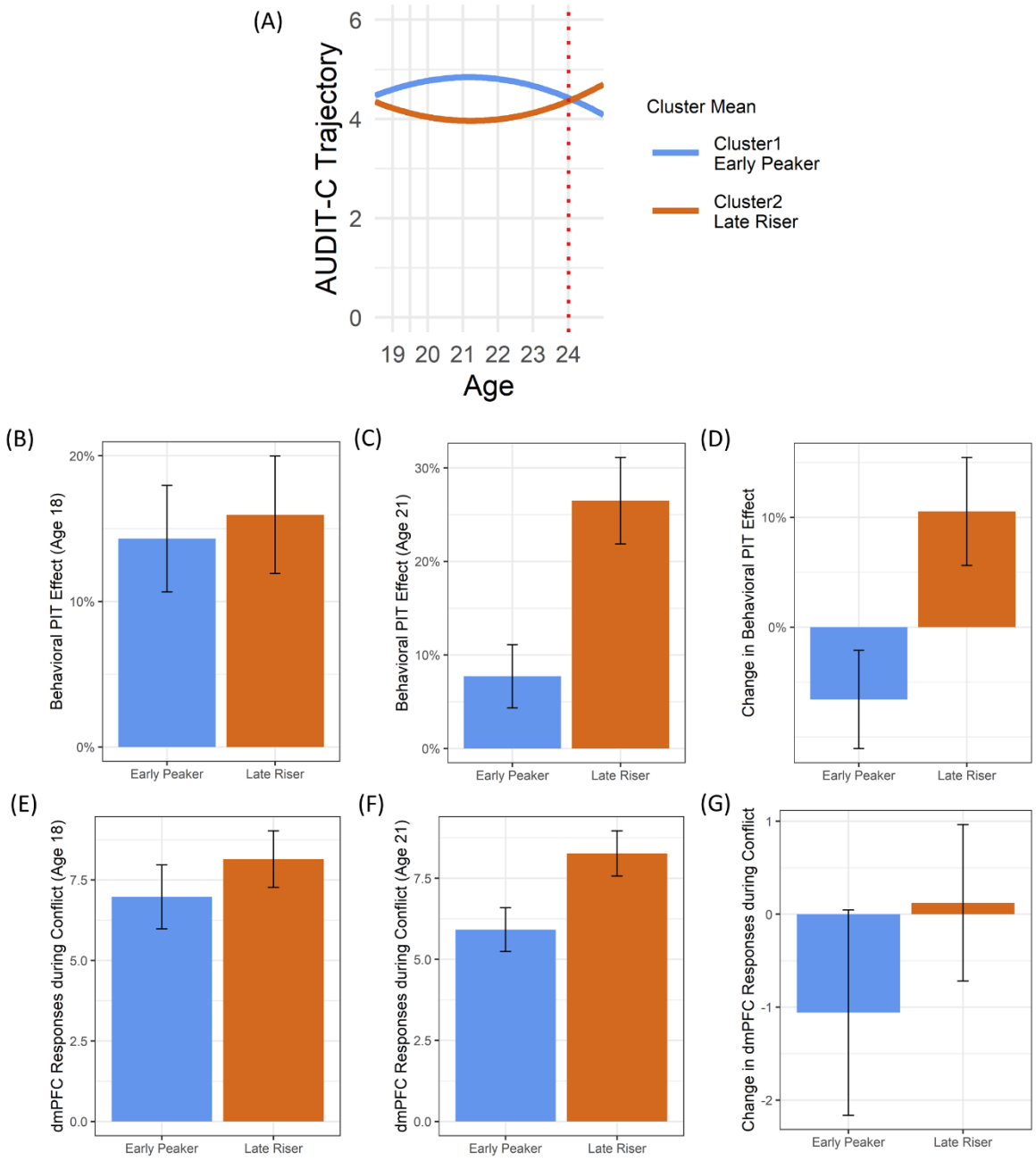


**Figure 21: Results of Alcohol Use Disorders Identification Test consumption score (AUDIT-C) clustering analysis**. (A) Trajectory of the two groups. Cluster 1 (mean starting point = 4.47; mean linear slope = 0.139; mean quadratic slope = -0.013) reached the peak around age 21 and was thus labeled as "early peaker" group. In contrast, cluster 2 (mean starting point = 4.35; mean linear slope = -0.142; mean quadratic slope = 0.013) first decreased and then developed prominently and labeled as the "late riser" group. (B) The two groups did not differ at age 18 with respect to the behavioral Pavlovian-to-instrumental transfer (PIT) effect. (C) The "late riser" group showed higher interference PIT effect at age 21 as compared to the "early

peaker" group. (D) The change in the interference PIT effect from age 18 to 21 was significantly different between the two groups. The "late riser" group also showed a change that was significantly different from zero, whereas the change was not different from zero for the "early peaker" group. (E) The two groups did not show different dorsomedial prefrontal cortex (dmPFC) responses during the conflict at age 18. (F) The dmPFC responses during conflict were stronger in the “later riser” group. (G) The changes in the dmPFC responses were not significantly different between the two groups.

### 4.4.8 Association Between Different Drinking Behaviors and Craving and Dependence

As shown in Table 7, the correlations between the OCDS and ADS and the AUDIT-C ranged from moderate to high at all available assessments and nominally increased over time. In contrast, the association between gram/occasion and OCDS was weak from ages 18 to 21, but this association was absent from ages 21 to 24. A similar pattern was found with the ADS sum score: the correlations with gram/occasion were moderate from ages 18 to 21 but attenuated after age 21.

*Table 6: fMRI results table*

| Incongruent vs. Congruent contrast in association with the behavioral PIT effect (ΔER) | | | | | | |
|---|---|---|---|---|---|---|
| Whole-brain results ($p_{uncorrected.} < .001$, cluster size ≥ 50) | | | | | | |
| Region | Side | Peak MNI | | | Peak-level *t* score | Cluster size |
| | | x | y | z | | |
| **Age 18 (*N* = 79)** | | | | | | |
| Supplementary motor area (including the lPFC and dmPFC clusters) | R | 8 | 18 | 48 | 5.53 | 8109 |
| Supramarginal gyrus | R | 54 | -44 | 38 | 5.46 | 1579 |
| Inferior parietal gyrus | L | -50 | -46 | 40 | 4.87 | 1035 |
| Middle frontal gyrus, orbital part | R | 36 | 44 | -10 | 4.50 | 106 |
| Inferior frontal gyrus, triangular part | L | -44 | 36 | 26 | 4.40 | 203 |
| Superior frontal gyrus, medial | R | 14 | 60 | 6 | 4.28 | 142 |
| Median cingulate and paracingulate gyri | R | 6 | -38 | 34 | 4.08 | 325 |
| Caudate (extended to ventral striatum) | R | 12 | 16 | 2 | 4.04 | 74 |
| Superior frontal gyrus, dorsolateral | L | -18 | 58 | 22 | 3.79 | 60 |
| Thalamus | L | -6 | -26 | -2 | 3.76 | 53 |
| Calcarine | R | 18 | -66 | 8 | 3.67 | 74 |
| Calcarine | L | -6 | -82 | 16 | 3.58 | 50 |
| Precuneus | L | -14 | -66 | 38 | 3.50 | 79 |
| **Age 21 (*N* = 79)** | | | | | | |
| Caudate (extended to ventral striatum) | R | 14 | 14 | 8 | 4.64 | 465 |
| Pallidum | L | -16 | 0 | 2 | 4.50 | 92 |
| Superior frontal gyrus, medial | L | -8 | 24 | 40 | 4.08 | 69 |
| Putamen | L | -24 | 16 | -6 | 3.78 | 78 |
| Middle frontal gyrus (including the lPFC cluster) | R | 38 | 52 | 10 | 3.73 | 73 |
| Anterior cingulate and paracingulate gyri | R | 6 | 44 | 18 | 3.62 | 61 |

*Table 7: Correlation between OCDS and ADS with different drinking measures*

| Obsessive Compulsive Drinking Scale (OCDS) | | | | | Alcohol Dependence Scale (ADS) | | | |
|---|---|---|---|---|---|---|---|---|
| OCDS and AUDIT-C | | | OCDS and gram/occasion | | ADS and AUDIT-C | | ADS and gram/occasion | |
| Age | spearman's rho | *p* | spearman's rho | *p* | spearman's rho | *p* | spearman's rho | *p* |
| 18 | --- | --- | 0.257 | .005 * | --- | --- | 0.442 | < .001 *** |
| 19 | 0.535 | < .001 *** | 0.266 | .016 * | 0.550 | < .001 *** | 0.409 | < .001 *** |
| 20 | 0.629 | < .001 *** | 0.330 | .004 * | 0.678 | < .001 *** | 0.413 | < .001 *** |
| 21 | 0.602 | < .001 *** | -0.001 | .992 | 0.593 | < .001 *** | 0.214 | .021 * |
| 22 | 0.666 | < .001 *** | 0.136 | .301 | 0.613 | < .001 *** | 0.250 | .054 |
| 23 | 0.600 | < .001 *** | 0.035 | .809 | 0.606 | < .001 *** | 0.240 | .089 |
| 24 | 0.750 | < .001 *** | 0.022 | .886 | 0.628 | < .001 *** | 0.218 | .150 |

*** *p* value < .001  * *p* value < .05

## 4.5 Discussion

In this study, we investigated the association between the interference PIT effect at ages 18 and 21 and drinking trajectories over 6 years until age 24 in a male community dwelling sample. The interference effect during PIT is behaviorally characterized by an increased ER during the conflict, i.e., when instrumental approach is required in the presence of a negatively-valenced Pavlovian cue or vice versa (instrumental avoidance required in the presence of a positively-valenced Pavlovian cue). At age 18, behavioral PIT effects (ΔER) were not significantly associated with drinking variables, however, a higher VS response during incongruent trials was associated with a higher baseline of the AUDIT-C and binge drinking score trajectories, but, contrary to what we hypothesized, a lower slope of the binge drinking score trajectory. Analyses of behavioral PIT data at age 21 indicated that a high interference effect predicted the increase of the AUDIT-C until age 24. This pattern was mirrored at the neural level: a stronger dmPFC response at age 21 was associated with an increase in the AUDIT-C over the next three years. Further cluster analysis with respect to the AUDIT-C trajectory revealed an "early peaker" group whose drinking behavior peaked already around age 21 and declined afterwards, and a "late riser" group whose drinking behavior started to develop prominently after age 21. Compared with the “early peakers”, the “late risers” showed not only a stronger behavioral interference PIT effect at age 21 but also a more pronounced increase of this effect from ages 18 to 21.

The results from the cluster analysis indicated that the interference PIT effect might point to an underlying mechanism driving the distinctive drinking patterns during young adulthood. But are there other variables associated with the different drinking patterns of the two groups? The profiles of the two groups may offer some insights. Specifically, the "early peakers" were found to experience more physical neglect during childhood and difficulties in describing their feelings. Previous studies supported the role of alexithymia in mediating the association between childhood trauma and alcohol addiction (Zdankiewicz-Ścigała & Ścigała, 2018; Zdankiewicz-Ścigała & Ścigała, 2020). Conversely, the "late risers" who developed prominently starting from age 21 had higher socioeconomic status and strong social motives when consuming alcohol. Although assessed on different levels, these findings indicate that a link may exist between environmental or psychosocial variables and cognitive measures like

PIT, in line with the recommendation to integrate socioeconomic and psychosocial aspects into the models of addiction (Heilig et al., 2016; Heinz et al., 2011; Hogarth, 2022).

Consistent with what we reported earlier in 139 participants with the baseline data (Chen et al., 2021d), we found that the VS responses during conflict were positively associated with the baseline of both the AUDIT-C and the binge drinking score trajectories in this subsample (*N* = 79). This supports the notion that the VS may play a central role during the initial bingeing and intoxication phase (Koob & Volkow, 2010). Contrarily, stronger functional VS activations during interference at age 18 were associated with more decrease or less increase in the binge drinking score over time. It is important to note that the statistical evidence for this association was weaker compared with the baseline associations; therefore, one needs to be cautious not to over-interpret this result. However, increased VS activation associated with less rather than more alcohol intake was also found with respect to alcohol cue exposure and alcohol PIT paradigms (Beck et al., 2012; Schad et al., 2019). If VS activation reflects attribution of salience to relevant cues (Heinz, 2002; Robinson & Berridge, 1993), it may under certain conditions contribute to behavior control.

Also, contrary to our hypothesis, we did not find any longitudinal association between the behavioral PIT effect and the binge drinking score trajectory. Why did we not detect such an association, given that the behavioral PIT effect at age 21 was associated with the AUDIT-C development? When examining the correlations between the linear slopes of the two drinking trajectories, we found that the individual AUDIT-C slopes were not significantly associated with the slopes of the binge drinking score trajectory. The low correlations indicate that the AUDIT-C develops differently and captures different information; indeed, alcohol intake may be high if frequently repeated, even if it is rather low per occasion. Through an exploratory analysis, we found that the AUDIT-C was highly correlated with alcohol craving and dependence throughout the six years, while the association between the per occasion drinking behavior and alcohol craving or dependence was only significant from ages 18 to 21 but became insignificant later. Therefore, in contrast to the AUDIT-C that evaluates both frequency and quantity of drinking, the sole amount of alcohol consumed during a typical occasion may not reflect craving or dependence during later stage of young adulthood.

Interestingly, stronger dmPFC responses during conflict at age 21 were associated with a more hazardous AUDIT-C trajectory (Figure 20B). Further cluster analysis confirmed this result—

"late risers" showed stronger dmPFC responses at age 21, which resembled the associations found with the behavioral PIT effect (Figure 21E-G). Previously, we found that stronger VS and weaker lPFC responses during conflict were associated with high-risk drinking and suspected that dmPFC might play similar roles as the lPFC (Chen et al., 2021d). However, the result here may suggest alternative functions of the dmPFC. On the one hand, it may encode a salience signal (Euston et al., 2012), and lower dmPFC responses might indicate that participants could focus attention towards stimuli relevant for the instrumental response and ignore distracting Pavlovian cues that interfere with the required instrumental behavior. In fact, stronger neural responses both in the medial prefrontal cortex and the VS have previously been found to be associated with enhanced motivation toward alcohol cues (Beck et al., 2012; Grüsser et al., 2004; Stuke et al., 2016). Alternatively, dmPFC responses may reflect error monitoring during conflict (Bastin et al., 2016). Investigating the role of dmPFC during the conflict between Pavlovian and instrumental control in more detail could help to address the role of this brain area in error monitoring during addiction development. Interestingly, low-risk drinkers showed stronger effective connectivity from the VS to the lPFC when dealing with the interference during PIT via a dynamic causal modelling approach (DCM) (Chen et al., 2021d). Network connectivity could be a critical factor for the development of drinking behavior during young adulthood (Veer et al., 2019).

The PIT predictors we included in our study had reliability ranging from 0.29 to 0.43, which could be considered as weak to moderate (Taylor, 1990), which may reflect the specific state-dependent components rather than stable traits. Our data indicate that such a state-dependent component may indeed exist, given that the change of the behavioral PIT effect predicted drinking trajectories, at least when we explored the differences between the two clusters (“early peakers” versus “late risers” in consumption). This is in line with changes in associated neurobiological systems including mesolimbic dopamine and cortical functions this development period (Flores-Barrera et al., 2014; Heng et al., 2011; Huppé-Gourgues & O'donnell, 2012). In accordance with the substantial changes in fronto-striatal circuits during this developmental period in late adolescence and early adulthood, the state component in the neural responses during the conflict was not consistently associated with the development of drinking behaviors.

Several limitations have to be addressed: first, we found different trajectory patterns (linear and quadratic slopes for AUDIT-C but linear trajectory for gram/occasion) to be optimal for the two trajectories. Since the binge drinking score and the AUDIT-C were assessed with different frequencies, we cannot rule out the possibility that this discrepancy may have happened since there were more AUDIT-C assessments available than the binge drinking score assessments, which allowed for more degrees of freedom in fitting a more complicated model to the individual-level drinking data. To test whether the AUDIT-C trajectory is different from binge drinking score trajectory, future studies should conduct more assessments within the same time interval.

Secondly, we used a fixed cluster number of two for the clustering analysis due to the limited sample size; future studies with a larger sample size could explore whether more subgroups with distinctive profiles could be identified.

Thirdly, on the behavioral level, only around two-thirds of the participants (62% at age 18 and 66% at age 21) showed a non-zero ER, which may have limited the power to predict the individual differences in the drinking trajectories. Future studies could improve the sensitivity of the measures to capture more subtle effects.

Lastly, we only included male participants, so these results cannot be generalized to non-male populations.

In summary, our six-year longitudinal study revealed that high error rates and their neural correlates due to conflict between Pavlovian and instrumental control can predict alcohol use trajectories. Through cluster analyses of the drinking trajectories, we identified two subgroups: the drinking behavior in the "late riser" group escalated after age 21, whereas the drinking of "early peakers" culminated at this age and then declined. The "late risers" showed enhanced dmPFC responses during conflict and three-times higher error rates during conflict between Pavlovian cues and instrumental responses in the PIT paradigm at age 21. Interestingly, this group also exhibited an increased behavioral PIT effect from age 18 to 21. Future studies could thus explore the dynamics of this interference PIT effect to predict risky drinking behaviors, potentially with more frequent PIT assessments. Such high-risk groups may then profit from targeted prevention and interventions.

## 4.6 Acknowledgements

This study was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft) (Grant Nos. 186318919 [FOR 1617], 178833530 [SFB 940], and 402170461 [TRR 265]). We thank Dr. Stephan Nebe for his contribution to the data collection.

# Chapter 5: General Discussion

## 5.1 Summary of Findings

The current thesis investigates whether the aberrant learning mechanisms—unbalanced goal-directed and habitual control and more substantial susceptibility to interference between Pavlovian cues and instrumental behaviors—predispose hazardous alcohol use during young adulthood. To achieve this, we followed up with a community sample of male social drinkers (*N* = 200) for 6 years from ages 18 to 24. Regarding the assessment of drinking behaviors, we focused on the AUDIT-C assessed every half a year starting 6 months after the baseline (at age 18.5) and binge drinking scores (gram/occasion), which were evaluated yearly.

For Study 1, we assessed a two-step sequential decision-making task during fMRI in this sample at age 18. We included the MB (goal-directed) and MF (habitual) control that we evaluated during this task in latent growth curve models to predict the participants' drinking trajectories from ages 18 to 21. We found that the MB behavioral control was negatively associated with the development of binge drinking score trajectories. Conversely, the MF RPE neural signals in the vmPFC and VS predisposed a higher starting point and a steeper increase or lower decrease in the consumption score trajectories over time, respectively. Overall, this finding demonstrates that unbalanced goal-directed and habitual controls could be regarded as an essential mechanism that influences risky alcohol use during young adulthood.

For Study 2, we examined the cross-sectional associations between the susceptibility to interference between Pavlovian cues and instrumental behavior regarding both behavioral and neural levels and risky (binge) drinking behavior at age 18. During the PIT task, the participants needed to "collect good shells" and "leave bad shells" during the presentation of appetitive or aversive Pavlovian cues. Conflicts could be elicited when the Pavlovian cues interfered with the ongoing instrumental behavior (incongruent condition; for example, when "collecting good shells" during the presentation of aversive Pavlovian cues). On the behavioral level, Pavlovian cues that were incongruent with instrumental behavior elicited higher ER than congruent cues did. The neural correlations with this interference PIT effect were found in the VS, lPFC, and dmPFC. We compared the interference PIT effect on behavioral and neural levels between the high- and low-risk drinkers, who drink more or less than 60 g of alcohol during a typical drinking occasion, respectively. The high-risk drinkers, in comparison to the

low-risk drinkers, exhibited more ER on the behavioral level. On the neural level, high-risk drinking was associated with stronger neural responses in the VS (on the trend level), weaker lPFC responses, and weaker connectivity the VS to the lPFC during conflicts, which indicates that bottom-up and top-down processes, as well as altered interplay within the brain network, all contributed to the more substantial susceptibility to conflicting Pavlovian cues.

After establishing the cross-sectional associations between the interference PIT effect and risky drinking behavior, we moved on to test whether the interference PIT effect works as a predisposing mechanism that could influence the development of risky trajectories. During Study 3, we aimed to predict the 6-year drinking trajectory from ages 18 to 24 with two PIT assessments, which were conducted at ages 18 and 21. The interference effect during PIT is characterized by an increased ER in the incongruent condition on the behavioral level; neural responses in the VS, lPFC, and dmPFC during the incongruent condition were extracted. Again, we constructed a binge drinking and a consumption score trajectory for drinking behaviors. On the behavioral level, we found that an increased interference PIT effect at age 21 was associated with risky drinking patterns in the consumption score trajectory, which started to emerge around this age and developed prominently until age 24. We also found two groups of people with distinctive drinking patterns through an exploratory clustering analysis: one "early peaker" group whose drinking behavior peaked around age 21 and then declined, and a "late riser" group whose drinking behavior started to develop prominently at age 21. Consistent with the described association, the "late risers", in comparison to the "early peakers", also displayed an increased behavioral PIT effect at age 21. Interestingly, their PIT effects also increased from ages 18 to 21, which suggests that the changes in their PIT effects also predisposed them to this at-risk drinking pattern. On the neural level, stronger neural responses in the VS during conflicts were associated with a higher baseline of the consumption and binge drinking score trajectories, which is consistent with Study 2. Contrary to what we hypothesized, stronger VS responses at age 18 were negatively associated with the development of the binge drinking trajectory; stronger dmPFC responses during conflicts at age 21 were associated with a riskier consumption score trajectory.

Taken altogether, the shift from goal-directed to habitual control, as informed by less MB and more MF control, appears to be a strong candidate for predisposing risky drinking trajectories during young adulthood. Therefore, several parameters—the MB behavioral score and the

RPE signals in the VS and vmPFC—could be considered promising (bio)markers for identifying vulnerable individuals and predicting risky alcohol use, especially during young adulthood. However, one should be cautiously optimistic since there is still ongoing debate about this experimental paradigm, modeling approach, and theoretical foundations, which I discuss in Section 5.2.

As for the interference PIT effect, our PIT paradigm, along with the analysis approach, has important theoretical implications: the transition from occasional to compulsive alcohol use may involve not only a shift from a specific to a general PIT effect (Hogarth et al., 2012) but also a more substantial interference PIT effect. It is thus crucial to examine the interaction between Pavlovian (appetitive and aversive cues) and instrumental (approach-avoidance) behaviors to assess the interference PIT effect. I discuss how one could integrate the interference PIT effect into the existing framework that explains the association between PIT processes and addictive behavior in Section 5.3.1.

Notably, the interference PIT effect and its changes from ages 18 to 21 on the behavioral level also seem to be promising predictors of future risky alcohol use development. On the neural level, the VS and dmPFC responses could also be considered biomarkers, although future studies are warranted to gain more insight into the role of dmPFC during the PIT processes. One limiting factor is that some of the participants did not make any errors in the incongruent condition; therefore, improving the sensitivity of this task (which is discussed in Section 5.3.2) to access the more subtle individual differences is essential for potential preventions and interventions. Additionally, all the PIT predictors had low to moderate reliability that can assess more state components than stable traits. Given that the behavioral PIT effect change is associated with the future development of risky drinking trajectories, one direction future researchers could pursue is exploring PIT dynamics in association with drinking behavior, potentially through more intensive sampling. The other direction is improving the reliability of the task to gain more psychometric properties to access the trait component. Sections 5.3.3 and 5.3.4 discuss these aspects.

## 5.2 The Debate Surrounding the Two-Step Task and How to Move Forward

Since Daw et al. (2011) proposed the two-step task, it has attracted enormous attention from researchers because it offers a promising way to experimentally assess the balance between

MF and MB behaviors. Moreover, it also provides a normative computational framework for explaining this balance. However, there have been heated discussions in the field in recent years. The critical discussions include whether the task measures what it is designed to measure, whether it has a solid theoretical foundation, whether the current computational framework could be further improved, and, ultimately, whether this dichotomous MB and MF control framework simplifies human behavior too much. This section provides a (non-exhaustive) literature review of the studies that address these issues and provides an outlook for future researchers.

### 5.2.1 MB Control Does Not Pay

Kool et al. (2016) argued that MB control should yield higher rewards when preferred over MF control since it demands more cognitive effort. Accordingly, the arbitration between MB and MF control depends on the cost-benefit analysis. MB control is preferred when it can demonstrate higher accuracy (in obtaining more benefits), considering its higher computational costs (Kool et al., 2017). Therefore, to elicit arbitration between MB and MF strategies, a task should pay off the higher cognitive demand by giving higher rewards to the MB system (Kool et al., 2018). Kool et al. (2016) demonstrated through simulations that higher MB control is not associated with a higher reward rate in the two-step task, which indicates that applying more demanding MB control does not pay off in this version of the task. Specifically, they found that many features of the task can lead to intact reward rates with more MB control, including slowly drifting reward probability that ranges from 0.25 to 0.75, probabilistic transition structure, ample action space during the second stage (four available actions), and the probabilistic nature of the second-stage reward. Based on these observations, Kool et al. (2016) proposed a new version of the two-step task that modifies all the mentioned features and found a significant positive association between the MB control ($\omega$) and the reward rate in this version.

Importantly, da Silva and Hare (2020) provided a story background and detailed instructions in the Daw version of the two-step task and found that their participants mainly applied MB control and little MF control. In their discussion, da Silva and Hare proposed that this finding may challenge the new version as it challenges the Daw version of the two-step task, since it is unlikely that the benefits of the MB control are different in the new version of the task in comparison to the Daw version. One exception is that providing more background

information for the transition structure could reduce the costs of cognitive efforts, but this would require further investigations. Therefore, this finding indicates that the MB control, as assessed in the two-step task, cannot evaluate a stable trait component as researchers thought. Moreover, da Silva and Hare (2020) found that participants could misunderstand the task differently but still behave like they were using a hybrid of MB and MF strategies. This study thus poses a question concerning whether behaviors during the two-step task could be mapped to the dichotomous MB and MF control as hypothesized, or rather, whether more strategies are involved.

### 5.2.2 Reconsidering the MF System

Shahar et al. (2019b) challenged the view that a MF system only assigns credit to the features of the two-step task that are relevant to the outcomes (monetary rewards) according to the experimental manipulation. They demonstrated that the MF system also assigns credit to low-level, outcome-irrelevant features such as the spatial-motor representation of cues, i.e., cue-to-key mappings. Notably, individuals who apply more MB control during a task assign less credit to any outcome-irrelevant spatial features. This study thus suggests that there could be a mismatch between the task manipulation and how the MF system actually represents the environment. Additionally, it was challenging to find a link between MF control and habitual responses that were made during the outcome-devaluation task, although a connection between MB control and goal-directed behavior was established during this task (Friedel et al., 2014; Gillan et al., 2015). However, it is important to note that researchers may need to overcome the barrier to inducing habits in laboratory settings (de Wit et al., 2018) to assess the association between habitual behavior and MF control more comprehensively.

Miller et al. (2019) re-evaluated the hypothesis that, in the RL framework, MF learning can be treated interchangeably with habits. Essentially, they pointed out the discrepancy between MF learning and the habits defined in the traditional associative learning framework. As described in the introduction, in the associative learning framework, habitual learning forms an association between stimulus and reward (S-R association) and is “value-free” (Dickinson & Balleine, 1994; Robbins & Everitt, 1999). In contrast, the MF system learns the association between stimulus and outcome, which is not “value-free.” Based on this argument, Miller et al. (2019) proposed a computational framework that can replace the current MF learning with a “value-free” component. In their model, the “value-free” habitual controller selects actions

based on the history of actions, rather than the outcome associated with the selected actions. Intriguingly, since habits could be regarded as a kind of "value-free" control in this framework, the MF and MB control could then be unified along a continuum axis that defines goal-directed behavior. It was also demonstrated that this alternative computational framework could explain the existing behavioral and neural data with a more solid theoretical foundation.

### 5.2.3 Better Models for the Two-Step Task

To predict the drinking trajectories, we applied the methods established in the original two-step paper (Daw et al., 2011). However, recent studies have suggested that this modeling approach could be improved in several ways. First, the reaction time in the second stage could be considered a parameter of interest since MB agents should react faster than MF ones after experiencing a common transition but should also react more slowly after a rare transition (Shahar et al., 2019a). Thus, including the reaction time in the model-agnostic measures and the MB measures (by combining reaction time with the drift diffusion model) could benefit internal and test-retest reliability.

Furthermore, we estimated the parameters of the RL model through the maximum likelihood methods. Recently, some researchers have suggested that the hierarchical estimation procedure produces more reliable estimates (Brown et al., 2020). However, Brown et al. also noted that the best-fitting model could be dependent on the sample—essentially, the model that worked best in one study cannot necessarily be generalized for other studies. Nevertheless, future researchers should consider the reliability of different measures when choosing an analysis strategy.

Future researchers could also consider a few alternative models when fitting computational models. First, researchers could use the alternative RL model in which MB and MF controls update the values sequentially (i.e., MB to inform the MF learning) to improve the model fit in comparison to the Daw model (Toyama et al., 2017). Additionally, the estimation of the balance between MB and MF controls could be biased when certain essential features during the learning process, such as forgetting and gradual preservation, are not included as parameters in the RL model (Toyama et al., 2019).

### 5.2.4 Is the Mind Dichotomous?

So far, promising future directions seem possible, such as investigating a modified task version (Kool et al., 2016) and testing alternative computational models (Shahar et al., 2019a; Toyama et al., 2017, 2019). However, the primary discussion still focuses on the theoretical view of a dichotomy between MB versus MF controls or goal-directed versus habitual controls. This dual-system theory is bound to be overly simplified (Daw, 2018) for a complex organ such as the brain. For instance, there could be an intermediate system between MB and MF controls, which some have named the successor representation system (Momennejad et al., 2017). The successor representation system caches long-term values in successor states. This algorithm has a lower degree of complexity, much like the MF learning, but preserves a certain level of flexibility, much like the MB control. Momennejad et al. (2017) also provided behavioral experimental evidence for the existence of such an intermediate algorithm.

Collins and Cockburn (2020) created an overview of the limitations of this dichotomy. First, MB and MF controls cannot be strictly segregated since they are both high-level computations that involve many lower-level ones. Furthermore, one cannot use the MB-MF dichotomy interchangeably with other dual-system process theories, including the goal-directed/habitual theory. Given these limitations, it is crucial to go beyond this over-simplified dichotomy to increase the dimensionality to investigate other strategies and computational methods to better understand learning and decision-making processes (Collins & Cockburn, 2020). Given that the “two types” frameworks are usually not testable and lack empirical support, going beyond the dual-process typology is essential for the cognitive science field in general (Melnikoff & Bargh, 2018).

### 5.2.5 How to Move Forward

To move forward, researchers first need to reconsider whether MF control could be regarded as habitual. If the answer is “no”, adjusting computational models or developing tasks to assess the arbitration between habitual and goal-directed behaviors would become essential. Given the difficulty of inducing habits in a laboratory setting (de Wit et al., 2018), developing reliable tasks to assess habits could be the first step. Recent studies suggested that it is important to consider the overlap between habits and motor skills, as well as how habits

strongly dominate action control (Du et al., 2022), as compared to goal-directed behaviors, within more familiar environment (Watson et al., 2022).

Second, the newly proposed task version (Kool et al., 2017) seems promising as a means of evaluating the balance between MB and MF control, although the task still assumes that the arbitration of MB and MF controls depends on a cost-benefit analysis. One alternative assumption, which seems promising, is that there is a "mixture of experts" (O'Doherty et al., 2021). Each expert uses their algorithm to produce a prediction that could also be mapped to neural circuits. A "controller" or "manager" tracks the reliability of each expert's prediction, most likely by computing a prediction error signal, and allocates more weight to those experts who make more reliable predictions and decides on the behaviors accordingly. Here, the cost of cognitive efforts does not need to be considered. Instead, according to the bias-variance trade-off perspective, the model with a sufficient degree of complexity is automatically preferred because simple models tend to have higher biases; the most complicated models are more vulnerable to errors when working with new input data. Additionally, the available cognitive resources naturally constrain the system.

This "mixture of experts" perspective also leads to the crucial view that it is essential to not restrict oneself in the MB vs. MF framework or any dual-process typology since this is an extremely challenging hypothesis to test. Alternatively, one should consider other well-known "experts" that have been left out of this framework, such as Pavlovian controls. To go beyond the dichotomous framework, researchers could investigate more forms of reinforcement learning on the algorithm level, such as hierarchical reinforcement learning (Botvinick et al., 2009), value-free reinforcement learning (Bennett et al., 2021), and the meta-reinforcement learning framework that is informed by the advances in artificial intelligence (Wang et al., 2018); the newly developed hierarchical Bayesian algorithms could also be considered (Schwöbel et al., 2021). In the end, I believe that a high degree of complexity is deemed necessary when one strives to understand high-level processes such as learning and decision-making.

## 5.3 Remarks on the PIT Task

### 5.3.1 How to Integrate the Single-Lever PIT Task Into the PIT Theory

Currently, there are two well-established types of PIT processes: the specific and general PIT. Typically, the two forms of PIT assess how different (appetitive) cues influence different types of approach behavior (Quail et al., 2017). However, this distinction does not consider the interaction between appetitive and aversive Pavlovian cues and instrumental approach-avoidance behaviors (e.g., the framework proposed by Guitart-Masip et al. [2014]). The following model may potentially incorporate our findings into the currently available theories.

Figure 22A presents the general PIT process. The instrumental response could either approach monetary rewards or avoid monetary losses. The influence of the instrumental system is depicted as orange stars on the axes; note that the influence can be different for approach and avoidance behaviors. In contrast, Pavlovian cues influence the ongoing instrumental behavior based on the associated outcomes: the cues promote approach tendencies when there is a potential reward but facilitate an avoidance tendency when there is possible loss. For some people, the effect of Pavlovian cues is more substantial, as illustrated by the dashed lines. Therefore, Pavlovian cues could hinder the desired instrumental responses when they are in conflict (i.e., when approach instrumental behavior is required when there is a potential loss or vice versa). Our studies revealed that more substantial susceptibility to Pavlovian cues (dashed vs. solid lines) when they conflict with the required instrumental behaviors could predispose an individual to risky alcohol use.

Different types of cues such as alcohol, chocolate, or juice (which are depicted with different colors) also impact instrumental behavior differently. For example, in this figure, the projection of alcohol cues on the approach-avoidance axis is more prominent than the projection of chocolate and juice, which suggests that this person may be particularly susceptible to alcohol cues. It is reasonable to assume that alcohol cues could elicit a more general motivating effect than other rewards and thus substantially impact the ongoing instrumental behavior.

Figure 22B represents a situation in which the instrumental behavior is to approach or avoid alcohol. Here, the influence of alcohol cues, in comparison to other cues, on instrumental behavior may be powerful, which could represent a specific PIT process.

This (preliminary and hypothetical) framework thus could integrate the seemingly different lines of research—the full transfer PIT task (e.g., van Timmeren et al., 2020) and the single-lever PIT task that we adopted in Studies 2 and 3—to better understand the roles of different processes in addiction. Future researchers could directly compare the full PIT task with the single-lever PIT task to find supporting evidence for this proposed framework or to refine it.

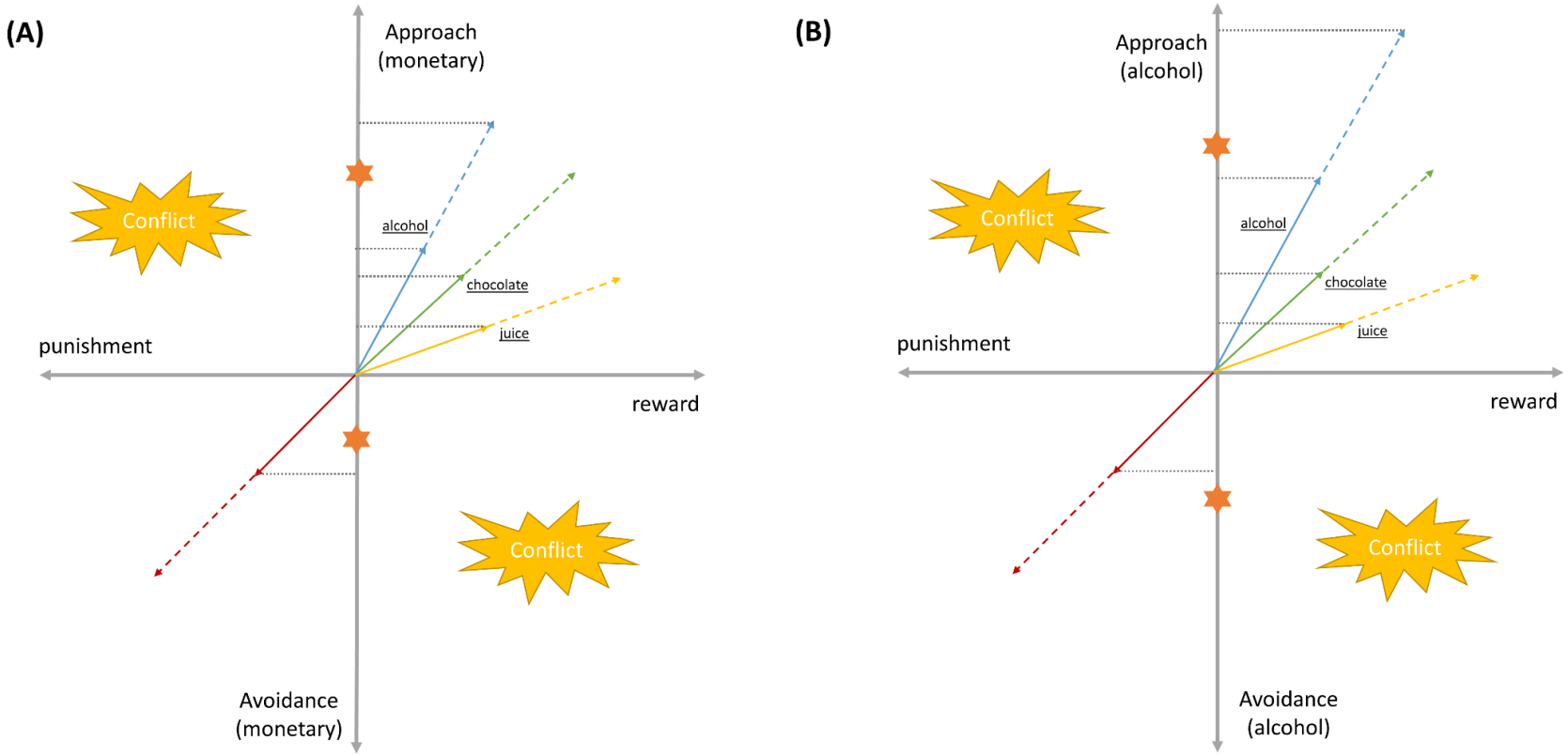


**Figure 22: Integrative Pavlovian-to-instrumental transfer framework** (further adaptations of Figure 1 in Guitart-Masip et al. [2014])

### 5.3.2 Improve the Sensitivity of the PIT Task

As measured by the increased ER in the incongruent condition, the interference effect did not follow a normal distribution; instead, many of the participants displayed a null or shallow PIT effect, which made it difficult to predict the drinking trajectories of these participants. To overcome this, we should better investigate the participants who displayed a null or shallow PIT effect to determine why they did not seem to be influenced by the Pavlovian cues during interference. One approach we could use is improving the paradigm to detect more subtle differences.

In Belanger et al. (under review), we tested an improved version of the PIT task. In this new version, the Pavlovian cues may have become more salient to the participants when a €10 reward or punishment was used instead of €1 or €2, which resulted in increased between-subject variances. Instead of button presses, this new version uses a joystick as the response device to capture the whole movement trajectory. Therefore, subtle reactions may be reflected in the movement trajectory or velocity. Moreover, in the current PIT experiment, the participants had to inhibit their responses to the "leave the shell." This feature of the task made it impossible to use the reaction time as one of the readouts of the experiment. Whereas in the joystick version of the task, the reaction time, including the start of the movement and the time it took for the participants to reach the threshold, could be used to characterize the PIT effect. Furthermore, when both the error rate and reaction time are available, one could consider fitting a drift-diffusion model (Forstmann et al., 2016; Ratcliff et al., 2016), which has been extended to account for value-based decision-making processes (Busemeyer et al., 2019). Fitting such a model could enhance reliability (Hedge et al., 2020) and allow researchers to understand the information accumulation processes during PIT better. We could thus ask questions such as, "Do Pavlovian and instrumental cues contribute differently to the different latent cognitive processes as indicated by the drift rate or the decision threshold, which subsequently led the participants to choose 'collecting' over 'leaving'?"

### 5.3.3 How to Exploit the State and Trait Components of PIT

During Study 3, the test-retest reliability of the interference PIT effect on the behavioral level was low. We suspect that the PIT may have reflected more on the state component than

assessing the trait component. Essentially, the state component seemed to be informative in our study since an increase in the behavioral PIT effect predisposed the consumption score's increase. Therefore, future researchers could further explore the dynamics of the PIT effect to prospectively predict drinking behavior. Ideally, the assessments could be conducted over a shorter period to better identify transition points that could inform predictions or interventions.

Conversely, researchers could also improve the reliability of the PIT task to better assess the trait component. The trait component is important, for example, in tracking individual development or determining the effects of interventions. Solutions could be sought at different levels. First, on the experiment design level, it is worth considering increasing the number of trials per condition and individual to have enough power to quantify individual differences (Rouder & Haaf, 2021). In one paper, the researchers propose that 100 trials per condition and person could be considered a rule of thumb for individual-level research. However, this also depends on how significant the trial-by-trial variance is. To optimize the trial and participant numbers quantitatively based on the between- and within-subject variance, one may check out the helpful online power contour tool (Baker et al., 2021a) that was provided by Baker et al. (2021b).

Second, from the perspective of the statistical approach, applying hierarchical models that account for the variations across the experimental trials in the statistical analyses has turned out to be helpful, as proven through the Stroop and Flanker tasks (Rouder & Haaf, 2019). Here, we calculated the individual mean interference PIT effect (ΔER) but did not consider the trial-by-trial variance on the individual level. One could apply a hierarchical model for the statistical analyses in the next step. Another proposal regarding the statistical analyses within hierarchical models is to pool the data across different experimental sessions when applying the mixed-effect model (Waltmann et al., 2021). When they analyzed the pooled data collected across two experimental sessions of the probabilistic reversal learning task, good to excellent test-retest reliability was achieved. Therefore, researchers could further apply a mixed model with the pooled data from the different PIT experimental sessions, as Waltmann et al. (2021) did, to investigate whether such approaches could improve reliability. Finally, fitting computational modeling may also improve reliability (Hedge et al., 2020; Hitchcock et

al., 2017). As mentioned in Section 5.3.3, a drift-diffusion model is also worth considering for the modified version of the PIT task.

### 5.3.4 Improving fMRI Reliability

Although the state component of the behavioral PIT effect seemed informative, the change in the neural PIT effect in the incongruent condition did not appear as informative since it was not associated with the further development of the drinking trajectories. Therefore, moving forward, future researchers should also consider improving the reliability of the PIT neural predictors.

Recently, researchers who conducted a large-scale meta-analysis (*N* = 1,008 across 90 experiments) demonstrated that the fMRI tasks generally have poor test-retest reliability (Elliott et al., 2020). Additionally, the neural responses within predefined ROIs across 11 commonly used fMRI tasks also demonstrated poor test-retest reliability. The researchers further argued that this poor test-retest reliability problem did not arise from the MRI measure itself but rather from the experimental method. Commonly adopted fMRI tasks have also been developed from robust group-level effects but not from the individual-difference perspective. Therefore, the first thing to do to improve the fMRI task reliability, I believe, is still to develop behavioral tasks that are robust to detect individual differences. For fMRI tasks, it is essential to further consider the stimuli that could maximize individual differences (Elliott et al., 2021; Elliott et al., 2020). Moreover, applying more naturalistic stimuli could keep participants engaged and potentially reduce the artifacts that can be caused by motions, attention, and fatigue (Elliott et al., 2021; Elliott et al., 2020). For the PIT task, future researchers could consider designing a PIT paradigm that is closer to a real-life situation, such as implementing more game-like instrumental tasks and presenting Pavlovian cues more naturally in the surrounding environment during the experiment.

Moreover, the fMRI analysis method could also be improved to provide better reliability. We extracted the neural responses from the incongruent condition within the three predefined ROIs and used these as individual predictors. Although this is typically done for the fMRI analysis, this method reduced the individual neural responses across the task to one regressor of interest and ignored the trial-by-trial or the voxel-by-voxel variability. As noted by Elliott et al. (2021), one way to improve reliability is to incorporate variability in the fMRI statistical

model (e.g., by adopting the latent variable approach) (Cooper et al., 2019). For instance, instead of averaging across one brain region, neural responses in different parcels (parcellation based on Gordon et al. [2016]) could be projected onto one latent variable when conducting an task fMRI analysis. In this way, the latent variable is free of measurement errors and reflects the actual variance shared by different parcels. Thus, it is more reliable and easier to replicate. This could be a promising direction for individual-difference-based fMRI analysis since this modeling approach explicitly accounts for the between-subject variance instead of treating such variances as errors (Cooper et al., 2019). Another method is modeling trial-by-trial variability in the hierarchical Bayesian framework (Chen et al., 2021a; Chen et al., 2021b).

Instead of focusing on an explanation, moving towards a predictive framework may emphasize individual differences during the analyses (Dubois & Adolphs, 2016; Yarkoni & Westfall, 2017). In this sense, applying machine learning methods such as multi-voxel pattern analysis (Norman et al., 2006) could be beneficial for reliability. Finally, once the data have been collected, the researchers would need to assess the test-retest reliability of neural responses under different contrasts and within different regions of interest and consider the reliable ones as the potential biomarkers (Fröhner et al., 2019).

## 5.4 How to Assess Drinking Behaviors to Capture the At-Risk State

Studies 1 and 3 characterized a consumption score and a binge-drinking score trajectory for the longitudinal prediction. We assumed that when the drinking trajectory displays an upward trend, it represents an intermediate state between occasional and compulsive alcohol use. This section discusses two aspects that are worth considering for assessing the at-risk state during future longitudinal studies.

### 5.4.1 Variables to Describe the At-Risk State

First, what are the variables of interest when describing the at-risk state, especially during young adulthood? It is important to note that this at-risk state could be different from diagnostic criteria, given that the level of dependence and alcohol-related harms are still evolving. We adopted consumption-related measures by describing two drinking trajectories: the binge drinking score (gram/occasion) and the consumption score (AUDIT-C) trajectories. Importantly, we found that the developmental patterns of the two drinking trajectories, as indicated by the slopes, did not display significant correlations. The consumption score

seemed to be strongly associated with dependence and craving, which typically represents a later stage in addiction development. This association may suggest that the consumption score could be more advantageous in identifying risky drinking patterns since it considers frequency measures and quantity.

However, the decision is more trivial than that. In Study 1, MB control was negatively associated with the binge drinking slope but not the consumption score trajectory. We suspected that this could be explained by a close link between binge drinking and deficits in executive function, as identified in previous studies (Carbia et al., 2018; Lannoy et al., 2019; Lees et al., 2019). A link between the consumption score and executive function seems to be missing in the literature. Future researchers should further investigate the relationship between the two variables from a longitudinal perspective. Unfortunately, our study was not suitable to answer such questions given that the consumption score and the binge drinking score trajectories were not sampled with the same frequency.

Importantly, some researchers have found that men and women between 18–24 respectively consume 10.1 and 8.1 drinks per binge drinking occasion on average (Naimi et al., 2010). High-intensity drinking—when people drink more than twice the binge drinking standard during one binge drinking occasion (Evans-Polce et al., 2017; Naimi et al., 2010; Patrick, 2016; Patrick et al., 2013; White et al., 2006)—has been proposed to describe this behavior. In comparison to binge drinkers, people who engage in high-intensity drinking are more likely to meet the AUD criteria, especially during early and young adulthood (Linden-Carmichael et al., 2017). Given the high prevalence of drinking during young adulthood that is closely linked to AUD, the high-intensity drinking trajectory could also be regarded as a proxy of the intermediate state that can develop into AUD. Future researchers could also investigate the associated mechanisms with this trajectory.

Ideally, with accumulating evidence, especially from longitudinal studies, researchers could better understand the distinctions between different assessments. Eventually, a compound measurement could be developed to best characterize the at-risk state and make developing targeted preventions more convenient.

### 5.4.2 Frequency of Assessments

What would be the optimal assessment frequency for a future longitudinal study to characterize the transition from occasional to compulsive alcohol use? It is essential to note that, when constructing the drinking trajectories, the time interval between assessments is assumed to reflect the underlying developing process (O'Rourke et al., 2021). This study demonstrates how drinking behavior develops within 6-month or yearly intervals, which may reflect changes on a macro time scale. Although such an approach has been broadly adopted, one aspect worth considering is whether the macro time scale assessments could capture the drinking behavior's underlying dynamics and critical stages.

One intriguing view is that AUD, in comparison to disorders that followed a well-defined developmental course (e.g., Parkinson's disease), is highly dynamic and dependent on states and environmental factors (Hitchcock et al., 2022). It is difficult to identify a universal developmental trajectory; instead, heterogeneous and highly dynamic individual courses can be observed. Therefore, it could be beneficial to conduct more assessments within shorter time intervals to capture the highly dynamic patterns, as Konova et al. (2020) did. Researchers could consider conducting smartphone-based studies to facilitate more dense sampling and to improve the efficacy of data collection (Gillan & Rutledge, 2021). Ideally, to better understand the dynamics of risky drinking behavior, one could consider a multiple-time-scale study design (Ram & Diehl, 2014), which collects data at different time scales. At the micro-level, bursting data could be collected across multiple days, which could be done efficiently with a smartphone-based design. At the macro level, longitudinal data could be collected with a slow time scale that takes measurements every 6 months or yearly. By collecting the behavioral data of interest at the same frequency, one may better understand the underlying mechanisms that drive the dynamic developmental process of risky drinking behavior and how this could develop into compulsive alcohol use.

## 5.5 Longitudinal Tools for Imaging Data Are Needed

In Studies 2 and 3, we were interested in predicting drinking trajectories with neural predictors. We only tested for the neural responses within the predefined ROIs in these analyses. Given the rich information in the fMRI data, one future direction is exploring the voxel-wise data and combining this flexibly with the LGCM models with the recently

developed neuropointillist toolbox (Madhyastha et al., 2018). Furthermore, in the LGCM models, we included different neural predictors from the ROIs in parallel to predict the drinking trajectories. This approach was well-justified in the two-step analysis; the computational model we adopted assumed no hierarchy between the MB and MF RPE signals on the neural level.

In contrast, this method could be one of the limitations of analyzing the PIT data since this parallel model did not capture the interplay between different brain regions. Specifically, the results from Study 2 suggest that processing the interference effects of Pavlovian cues relies on the corticostriatal circuit. The subcortical structures such as VS communicate intensively with the cortical part, such as the lPFC and dmPFC, concerning response selection and cognitive control (Haber, 2016; Peters et al., 2016). In this case, the connectivity between the ROIs is of particular interest since one could consider the interplay between different regions and the hierarchy within the network by assuming the forward and backward connectivity. For Study 2, we investigated the interplay between the cortical and subcortical areas through effective connectivity with the DCM analysis. However, this posed some difficulties for the longitudinal analysis given the already complicated model structures with two PIT assessments.

To analyze how the effective connectivity from two MRI sessions could predict the drinking trajectories, researchers may need to develop a DCM framework that allows the input of two or even more MRI sessions while accounting for both inter- and intra-individual variances. One could extract the slopes calculated with the LGCM model to look for an association with the drinking data. Including the individual slopes in the DCM model should be non-trivial given the recent development in the Parametric Empirical Bayes method that could be applied to test for individual or group differences in effective connectivity (Zeidman et al., 2019a; Zeidman et al., 2019b). Alternatively, since the DCM is implemented within a Bayesian framework to make the LGCM framework more compatible with the DCM, future researchers could also consider fitting a growth curve within the Bayesian framework (Oravecz & Muth, 2018), which might facilitate the combination of the different models.

## 5.6 A Unified Framework of Pavlovian, Habitual, and Goal-Directed Controls

Finally, I would like to discuss how to connect different learning mechanisms. Until this point in the thesis, it may seem like one needed to develop various experimental tasks and follow different lines of thought when studying the interaction between goal-directed behavior and habitual or Pavlovian learning. Indeed, as outlined in the introduction, the two lines of research followed quite different developmental paths, especially from the experimental design perspective. Pavlovian-to-instrumental tasks were originally developed for animal studies and were later extended to human studies (Cartoni et al., 2016; Holmes et al., 2010). The current version of the two-step task was developed using inspiration from the reinforcement learning framework (Daw et al., 2011; Daw et al., 2005).

Some researchers have proposed that the outcome-specific PIT may have some goal-directed properties, although there is still ongoing debate (Mahlberg et al., 2021). Conversely, drug cues could also enhance habitual behaviors (Everitt & Robbins, 2016), possibly through a general PIT process. As for the transition from occasional to compulsive alcohol use, while drug cues can acquire some general motivating properties (i.e., shifting from specific to general PIT processes as indicated in Hogarth et al. [2012]), they can also enhance habitual behaviors (Everitt & Robbins, 2016). Empirically, with the same two tasks that we used, Sebold et al. found that the stronger influence of Pavlovian cues on instrumental behavior was associated with decreased MB control (Sebold et al., 2016). Therefore, there is no reason why the two concepts should remain segregated on the theoretical and empirical levels while moving forward with the addiction research.

One possible way to combine and manipulate the two types of behavioral controls could be within the reinforcement learning framework. The two-step task is an example of implementing both goal-directed and habitual controls in the reinforcement learning framework. The influence of Pavlovian control on instrumental behavior during learning has also been assessed with the reinforcement learning model with a valenced go/no-go task (Guitart-Masip et al., 2012). Therefore, current computational models seem to be able to integrate all these different compartments from the long-standing associative learning framework into computational models within a normative framework.

Future researchers could also consider integrating the reward/loss manipulation of the valenced go-no-go task into the two-step task, or better, the improved version or variants of the two-step task from an experimental design perspective. Ideally, one would manipulate Pavlovian, habitual, and goal-directed controls within one experimental setting. In addition to developing experimental paradigms, progress in the computational framework is also needed; for example, recently, a Bayesian framework was developed to describe the arbitration between Pavlovian and instrumental actions through the valenced go-no-go task (Dorfman & Gershman, 2019; Gershman et al., 2021). When combining the different concepts into one unified normative framework, such a computational model should also be considered.

Taken all together, I think that the Pavlovian-to-instrumental transfer and the goal-directed and habitual controls could be united as crucial concepts to explain risky drinking and the transition towards compulsive alcohol use. Once better experimental tasks that allow for the manipulation of different processes are developed, and a normative computational framework that could incorporate different concepts is achieved, we can better understand learning as an important mechanism that influences the development of addiction. Beyond addiction, a more profound understanding of the learning mechanisms could also be the key to understanding human cognition.

# Appendix

## A Supplementary Materials: Study 1

### A.1 Recruitment Procedure and Inclusion Criteria

201 18-year-old male participants recruited via the local registration offices in Berlin and Dresden completed the baseline assessments (more details see [Nebe et al., 2018]). The inclusion criteria stated participants had to have partaken in at least two drinking occasions within the last three months, normal or corrected-to-normal vision, right-handedness, no contraindications for MRI scanning, neither a history of nor current diagnosis of a mental disorder, and no substance dependence (excluding nicotine). Participants who fulfilled the alcohol abuse criteria were also included.

### A.2 Construction and Descriptive Statistics of Drinking Trajectories (Consumption Score and Binge Drinking Score)

The average alcohol intake per drinking occasion (gram/occasion) during the past year from the M-CIDI interview assesses the alcohol drinking behavior on a typical drinking occasion. However, instead of separating bingers and non-bingers based on a cut-off, the measure of gram/occasion was treated as a continuous variable to preserve more information; this also follows the line of thought in moving towards a dimensional approach in characterizing drinking behavior as done in DSM-V (Hasin et al., 2013). The AUDIT consumption score (AUDIT-C) was also used to construct drinking trajectories, which is comprised of the sum of the first three items of the 10-item AUDIT questionnaire. It has been suggested to be a very sensitive measure of risky drinking (Dawson, 2011) but it is not limited to binge drinking. Since the AUDIT-C has proven to be nearly as effective as the full 10-question AUDIT (Dawson, 2011) and a quarter of the participants in our sample reported increased AUDIT-C that was not detected with the total score (Kuitunen-Paul et al., 2018) , we built the drinking trajectory based on the AUDIT-C.

*Table S1: Descriptive statistics for AUDIT consumption score and binge drinking score*

| AUDIT consumption score (AUDIT-C) | | | | | | |
|---|---|---|---|---|---|---|
| | N (non-missing rate of *N* = 133)* | Min | Median | Max | Mean | SD |
| FU 06 | 71 (53.4%) | 0 | 4 | 9 | 4.46 | 2.02 |
| FU 12 | 89 (66.9%) | 0 | 4 | 9 | 4.18 | 2.05 |
| FU 18 | 97 (72.9%) | 0 | 4 | 9 | 4.48 | 2.08 |
| FU 24 | 98 (73.7%) | 0 | 4 | 9 | 4.34 | 2.07 |
| FU 30 | 94 (70.7%) | 0 | 4 | 9 | 4.17 | 1.97 |
| FU 36 | 99 (74.4%) | 0 | 4 | 10 | 4.09 | 1.96 |
| Binge drinking score (gram alcohol / drinking occasion) past year | | | | | | |
| | N (non-missing rate of *N* = 146) | Min | Median | Max | Mean | SD |
| BL | 146 (100%) | 18 | 54 | 225 | 70.95 | 44.19 |
| FU 12 | 118 (80.8%) | 0 | 45 | 252.9 | 59.99 | 41.78 |
| FU 24 | 95 (65.1%) | 0 | 45 | 157.5 | 56.85 | 37.88 |
| FU 36 | 99 (67.8%) | 0 | 36 | 153 | 43.64 | 34.10 |

** Thirteen people were excluded for the AUDIT consumption analyses due to the lack of valid AUDIT assessments over the three years.*

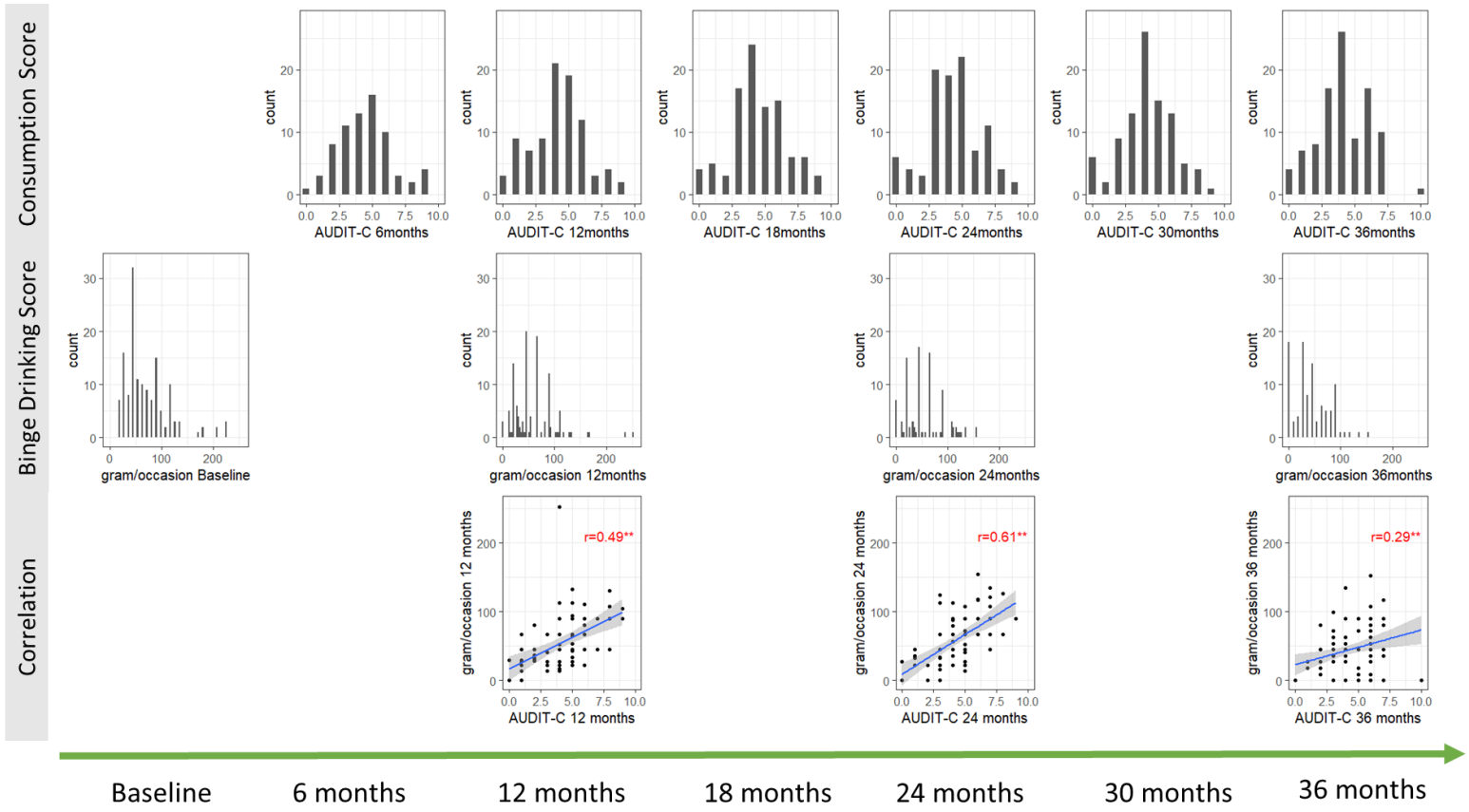


**Figure S1: Histograms of the drinking behaviors during the past year.** Alcohol Use Disorders Identification Test consumption score (AUDIT-C) started from six months after the baseline and was measured every six months; binge drinking score (gram/occasion) was measured every year from baseline to the third-year follow-up. The third row showed the correlation between the two variables whenever assessed at the same time point; the correlation became weaker at the follow-up 36 months.

### A.3 Cognitive Ability Assessment

At the baseline assessment, three aspects of cognitive ability were assessed: processing speed, working memory, and crystallized intelligence. In the Digit Symbol Substitution Task (DSST) (Wechsler, 1997) assessing the processing speed, subjects needed to substitute numbers with their corresponding abstract symbols according to a list. The number of successful substitutions within 120 seconds was used as the measure of processing speed. The Trail Making Test (TMT) also assessed the processing speed (Reitan, 1979). The task consisted of two parts. In part A, Participants were instructed to connect the 25 numbers (1-25) in ascending orders. In part B, both numbers (1-13) and letters (A-L) were presented; participants needed to alternate the numbers and letters when connecting (i.e., 1-A-2-B-3-C…). The time (in seconds) it took to complete the task was the indicator of task performance. Working memory capacity was assessed via the Digit Span Backwards Test (DSbw; Wechsler, 1997), which was indicated by the maximum number of digits a participant could recall in reverse order. In the German vocabulary test (Mehrfachwahl-Wortschatz-Intelligenztest, MWT-B [Lehrl, 2005]) assessing crystallized intelligence, the participants identified an actual word among a word list with four more nonsense words (Lehrl, 2005). The performance was the number of correct answers out 37-word lists.

We correlated tests all six two-step predictors with the three cognitive ability variables (see Table S2). The Spearman's rho was applied for the correlation tests due to the non-normality of the cognitive function measures. Among all the two-step predictors, only the MB behavioral score was associated with the cognitive functions: More MB behavioral control was associated with higher processing speed as indicated by more successful substitutions in DSST and less time used for the TMT-B; more MB behavioral control was also associated with better working memory capacity assessed by the digit span task.

*Table S2 : Correlation between the two-step predictors and the cognitive functions*

| ***Cognitive function*** | ***Two-step predictors*** | ***Spearman's rho*** | ***p*** | |
|---|---|---|---|---|
| *TMT-A (second)* | MF score | 0.017 | .839 | |
| | MB score | -0.085 | .307 | |
| | MB VS | 0.03 | .717 | |
| | MB vmPFC | 0.132 | .113 | |
| | MF VS | -0.057 | .498 | |
| | MF vmPFC | -0.071 | .393 | |
| *TMT-B (second)* | MF score | 0.039 | .639 | |
| | **MB score** | **-0.178** | **.032** | * |
| | MB VS | 0.055 | .509 | |
| | MB vmPFC | 0.126 | .130 | |
| | MF VS | -0.028 | .740 | |
| | MF vmPFC | -0.041 | .624 | |
| *DSbw (maximum digits)* | MF score | -0.065 | .433 | |
| | **MB score** | **0.176** | **.033** | * |
| | MB VS | -0.014 | .871 | |
| | MB vmPFC | -0.003 | .971 | |
| | MF VS | -0.056 | .505 | |
| | MF vmPFC | -0.023 | .782 | |
| *DSST* | MF score | -0.027 | .747 | |
| *(successful substitutions)* | **MB score** | **0.24** | **.003** | ** |
| | MB VS | 0.011 | .893 | |
| | MB vmPFC | 0.04 | .627 | |
| | MF VS | -0.103 | .216 | |
| | MF vmPFC | -0.014 | .865 | |
| *MWT-B (correct answers)* | MF score | -0.027 | .744 | |
| | MB score | 0.143 | .086 | |
| | MB VS | -0.027 | .748 | |
| | MB vmPFC | 0.075 | .370 | |
| | MF VS | 0.101 | .226 | |
| | MF vmPFC | 0.074 | .376 | |

* $p < .05$, ** $p < .01$; $N = 146$

Abbreviations: model-based (MB); model-free (MF); ventromedial prefrontal cortex (vmPFC); ventral striatum (VS)

## A.4 Descriptive Statistics of Other Assessed Variables Over the Three Years

Table S3: Descriptive statistics of other assessed variables over the three years

| *Measures* | *Time Point ** | *Valid values* | *Missing rate* | *Mean* | *Std* | *Min* | *1st quartile* | *Median* | *3rd quartile* | *Max* |
|---|---|---|---|---|---|---|---|---|---|---|
| *Age* | **BL** | 146 | 0.00% | 18.35 | 0.19 | 18.05 | 18.21 | 18.31 | 18.45 | 18.83 |
| | **FU06** | 115 | 21.23% | 18.98 | 0.27 | 18.61 | 18.79 | 18.90 | 19.12 | 20.25 |
| | **FU12** | 129 | 11.64% | 19.41 | 0.20 | 19.06 | 19.25 | 19.37 | 19.53 | 20.02 |
| | **FU18** | 98 | 32.88% | 19.95 | 0.21 | 19.51 | 19.79 | 19.90 | 20.08 | 20.44 |
| | **FU24** | 120 | 17.81% | 20.41 | 0.21 | 20.08 | 20.24 | 20.36 | 20.55 | 21.06 |
| | **FU30** | 78 | 46.58% | 20.93 | 0.24 | 20.57 | 20.74 | 20.89 | 21.05 | 22.09 |
| | **FU36** | 99 | 32.19% | 21.44 | 0.24 | 21.05 | 21.28 | 21.38 | 21.55 | 22.29 |
| *Age - 1st Bingeing* | **BL** | 100 | 31.51% | 16.55 | 0.85 | 14.00 | 16.00 | 16.50 | 17.00 | 18.22 |
| *Age - 1st Drink* | **BL** | 146 | 0.00% | 14.34 | 1.38 | 10.00 | 14.00 | 14.00 | 15.00 | 17.92 |
| *Age - 1st time drunken* | **BL** | 139 | 4.79% | 15.80 | 1.14 | 12.00 | 15.00 | 16.00 | 17.00 | 18.33 |
| *Drinking Frequency past-year* | **BL** | 146 | 0.00% | 2.38 | 0.69 | 1.00 | 2.00 | 2.00 | 3.00 | 4.00 |
| | **FU12** | 119 | 18.49% | 2.49 | 0.81 | 0.00 | 2.00 | 3.00 | 3.00 | 4.00 |
| | **FU24** | 95 | 34.93% | 2.47 | 0.97 | 0.00 | 2.00 | 3.00 | 3.00 | 5.00 |
| | **FU36** | 99 | 32.19% | 2.53 | 0.97 | 1.00 | 2.00 | 3.00 | 3.00 | 5.00 |
| *Drinking per day past-year (in gram pure ethanol)* | **BL** | 146 | 0.00% | 11.25 | 13.37 | 0.64 | 3.21 | 7.07 | 15.43 | 112.50 |
| | **FU12** | 118 | 19.18% | 10.32 | 10.28 | 0.00 | 3.21 | 6.43 | 14.46 | 66.15 |
| | **FU24** | 95 | 34.93% | 10.93 | 12.27 | 0.00 | 3.05 | 6.43 | 14.46 | 67.50 |
| | **FU36** | 99 | 32.19% | 9.45 | 10.11 | 0.00 | 2.44 | 6.36 | 12.44 | 40.50 |
| *Drinking per occasion past-year (in gram pure ethanol)* | **BL** | 146 | 0.00% | 70.95 | 44.19 | 18.00 | 45.00 | 54.00 | 90.00 | 225.00 |
| | **FU12** | 118 | 19.18% | 60.00 | 41.78 | 0.00 | 27.45 | 45.00 | 81.00 | 252.90 |
| | **FU24** | 95 | 34.93% | 56.85 | 37.88 | 0.00 | 22.50 | 45.00 | 83.70 | 157.50 |
| | **FU36** | 99 | 32.19% | 43.64 | 34.10 | 0.00 | 22.50 | 36.00 | 67.50 | 153.00 |
| *AUDIT-C (consumption score)* | **FU06** | 71 | 51.37% | 4.47 | 2.02 | 0.00 | 3.00 | 4.00 | 6.00 | 9.00 |
| | **FU12** | 89 | 39.04% | 4.18 | 2.05 | 0.00 | 3.00 | 4.00 | 5.00 | 9.00 |
| | **FU18** | 97 | 33.56% | 4.49 | 2.08 | 0.00 | 3.00 | 4.00 | 6.00 | 9.00 |
| | **FU24** | 98 | 32.88% | 4.34 | 2.07 | 0.00 | 3.00 | 4.00 | 5.00 | 9.00 |
| | **FU30** | 94 | 35.62% | 4.17 | 1.97 | 0.00 | 3.00 | 4.00 | 5.00 | 9.00 |
| | **FU36** | 99 | 32.19% | 4.09 | 1.96 | 0.00 | 3.00 | 4.00 | 6.00 | 10.00 |
| *AUDIT total score* | **FU06** | 71 | 51.37% | 6.00 | 4.17 | 0.00 | 4.00 | 5.00 | 7.50 | 25.00 |
| | **FU12** | 89 | 39.04% | 5.97 | 3.92 | 0.00 | 3.00 | 5.00 | 8.00 | 18.00 |
| | **FU18** | 97 | 33.56% | 6.28 | 4.14 | 0.00 | 4.00 | 6.00 | 8.00 | 25.00 |
| | **FU24** | 98 | 32.88% | 5.98 | 4.00 | 0.00 | 4.00 | 5.00 | 7.00 | 23.00 |
| | **FU30** | 94 | 35.62% | 5.81 | 3.92 | 0.00 | 3.00 | 5.00 | 7.00 | 22.00 |
| | **FU36** | 99 | 32.19% | 5.46 | 3.33 | 0.00 | 3.00 | 5.00 | 7.00 | 16.00 |

| *Measures* | *Time Point ** | *Valid values* | *Missing rate* | *Mean* | *Std* | *Min* | *1st quartile* | *Median* | *3rd quartile* | *Max* |
|---|---|---|---|---|---|---|---|---|---|---|
| *OCDS total*[1] | **BL** | 142 | 2.74% | 3.51 | 3.04 | 0.00 | 1.00 | 2.00 | 5.00 | 18.00 |
| | **FU12** | 101 | 30.82% | 2.63 | 2.59 | 0.00 | 1.00 | 2.00 | 4.00 | 11.00 |
| | **FU24** | 105 | 28.08% | 2.74 | 2.45 | 0.00 | 1.00 | 2.00 | 4.00 | 10.00 |
| | **FU36** | 98 | 32.88% | 2.77 | 2.82 | 0.00 | 1.00 | 2.00 | 3.00 | 15.00 |
| *ADS Score*[2] | **BL** | 141 | 3.42% | 4.69 | 4.08 | 0.00 | 2.00 | 4.00 | 7.00 | 30.00 |
| | **FU12** | 103 | 29.45% | 3.47 | 3.11 | 0.00 | 1.00 | 3.00 | 5.00 | 18.00 |
| | **FU24** | 107 | 26.71% | 3.06 | 2.83 | 0.00 | 1.00 | 3.00 | 5.00 | 14.00 |
| | **FU36** | 99 | 32.19% | 3.21 | 3.09 | 0.00 | 1.00 | 2.00 | 5.00 | 14.00 |
| *BIS attention*[3] | **BL** | 145 | 0.68% | 8.80 | 1.98 | 5.00 | 7.00 | 9.00 | 10.00 | 14.00 |
| | **FU12** | 101 | 30.82% | 8.60 | 2.05 | 5.00 | 7.00 | 9.00 | 10.00 | 14.00 |
| | **FU24** | 105 | 28.08% | 8.59 | 2.03 | 5.00 | 7.00 | 8.00 | 10.00 | 14.00 |
| | **FU36** | 99 | 32.19% | 8.49 | 1.92 | 5.00 | 7.00 | 8.00 | 10.00 | 15.00 |
| *BIS motor*[3] | **BL** | 145 | 0.68% | 10.05 | 2.38 | 5.00 | 8.00 | 10.00 | 12.00 | 18.00 |
| | **FU12** | 101 | 30.82% | 10.13 | 2.39 | 6.00 | 8.00 | 10.00 | 12.00 | 17.00 |
| | **FU24** | 105 | 28.08% | 10.01 | 2.48 | 6.00 | 8.00 | 10.00 | 11.00 | 18.00 |
| | **FU36** | 99 | 32.19% | 10.02 | 2.54 | 6.00 | 8.00 | 10.00 | 12.00 | 17.00 |
| *BIS non-planning*[3] | **BL** | 145 | 0.68% | 11.08 | 2.75 | 5.00 | 9.00 | 11.00 | 13.00 | 19.00 |
| | **FU12** | 101 | 30.82% | 10.55 | 3.36 | 5.00 | 8.00 | 10.00 | 13.00 | 19.00 |
| | **FU24** | 105 | 28.08% | 10.36 | 2.92 | 5.00 | 9.00 | 10.00 | 12.00 | 17.00 |
| | **FU36** | 99 | 32.19% | 10.12 | 3.03 | 5.00 | 8.00 | 10.00 | 12.00 | 17.00 |
| *BIS sum score*[3] | **BL** | 145 | 0.68% | 29.93 | 5.24 | 18.00 | 27.00 | 30.00 | 33.00 | 45.00 |
| | **FU12** | 101 | 30.82% | 29.29 | 6.04 | 18.00 | 25.00 | 29.00 | 33.00 | 44.00 |
| | **FU24** | 105 | 28.08% | 28.96 | 5.66 | 18.00 | 25.00 | 29.00 | 32.00 | 45.00 |
| | **FU36** | 99 | 32.19% | 28.63 | 5.12 | 17.00 | 26.00 | 29.00 | 32.00 | 40.00 |
| *DMQ Conformity*[4] | **BL** | 143 | 2.05% | 6.06 | 1.81 | 5.00 | 5.00 | 5.00 | 6.00 | 16.00 |
| | **FU12** | 94 | 35.62% | 5.93 | 1.82 | 5.00 | 5.00 | 5.00 | 6.00 | 14.00 |
| | **FU24** | 98 | 32.88% | 6.03 | 1.44 | 5.00 | 5.00 | 6.00 | 6.00 | 11.00 |
| | **FU36** | 94 | 35.62% | 6.15 | 1.93 | 5.00 | 5.00 | 5.00 | 6.00 | 16.00 |
| *DMQ Coping*[4] | **BL** | 143 | 2.05% | 6.85 | 2.63 | 5.00 | 5.00 | 6.00 | 7.00 | 18.00 |
| | **FU12** | 94 | 35.62% | 6.29 | 1.71 | 5.00 | 5.00 | 6.00 | 7.00 | 13.00 |
| | **FU24** | 98 | 32.88% | 6.83 | 2.75 | 5.00 | 5.00 | 6.00 | 7.75 | 21.00 |
| | **FU36** | 94 | 35.62% | 6.16 | 1.71 | 5.00 | 5.00 | 5.00 | 7.00 | 13.00 |
| *DMQ Enhancement*[4] | **BL** | 143 | 2.05% | 11.79 | 4.72 | 5.00 | 8.00 | 11.00 | 15.00 | 25.00 |
| | **FU12** | 94 | 35.62% | 10.93 | 4.66 | 5.00 | 7.00 | 10.00 | 14.00 | 23.00 |
| | **FU24** | 98 | 32.88% | 11.53 | 5.07 | 5.00 | 7.00 | 11.00 | 14.00 | 25.00 |
| | **FU36** | 94 | 35.62% | 11.49 | 5.20 | 5.00 | 7.00 | 10.00 | 15.75 | 23.00 |
| *DMQ Social*[4] | **BL** | 143 | 2.05% | 13.58 | 4.43 | 5.00 | 10.00 | 14.00 | 17.00 | 23.00 |
| | **FU12** | 94 | 35.62% | 11.30 | 4.18 | 5.00 | 9.00 | 11.00 | 14.00 | 23.00 |
| | **FU24** | 98 | 32.88% | 12.93 | 4.70 | 5.00 | 10.00 | 12.00 | 16.75 | 23.00 |
| | **FU36** | 94 | 35.62% | 12.95 | 4.53 | 5.00 | 10.00 | 12.00 | 17.00 | 24.00 |

| *Measures* | *Time Point ** | *Valid values* | *Missing rate* | *Mean* | *Std* | *Min* | *1st quartile* | *Median* | *3rd quartile* | *Max* |
|---|---|---|---|---|---|---|---|---|---|---|
| ***AEQ Total[5]*** | **BL** | 143 | 2.05% | 29.02 | 4.52 | 19.00 | 26.50 | 29.00 | 33.00 | 38.00 |
| | **FU12** | 101 | 30.82% | 28.72 | 4.68 | 19.00 | 26.00 | 29.00 | 32.00 | 38.00 |
| | **FU24** | 107 | 26.71% | 27.94 | 4.67 | 19.00 | 25.00 | 28.00 | 31.00 | 38.00 |
| | **FU36** | 99 | 32.19% | 27.83 | 4.85 | 19.00 | 24.00 | 28.00 | 31.00 | 37.00 |
| ***SURPS Anxiety sensitivity[6]*** | **BL** | 145 | 0.68% | 10.54 | 2.43 | 5.00 | 9.00 | 11.00 | 13.00 | 17.00 |
| | **FU12** | 101 | 30.82% | 11.37 | 2.36 | 5.00 | 10.00 | 12.00 | 13.00 | 16.00 |
| | **FU24** | 105 | 28.08% | 11.33 | 2.76 | 6.00 | 10.00 | 12.00 | 14.00 | 17.00 |
| | **FU36** | 99 | 32.19% | 10.28 | 2.68 | 5.00 | 8.00 | 10.00 | 12.00 | 16.00 |
| ***SURPS Hopelessness[6]*** | **BL** | 145 | 0.68% | 11.95 | 2.80 | 7.00 | 10.00 | 12.00 | 14.00 | 23.00 |
| | **FU12** | 101 | 30.82% | 11.34 | 3.36 | 7.00 | 9.00 | 11.00 | 14.00 | 25.00 |
| | **FU24** | 105 | 28.08% | 11.88 | 3.09 | 7.00 | 9.00 | 12.00 | 14.00 | 22.00 |
| | **FU36** | 99 | 32.19% | 10.74 | 2.88 | 7.00 | 8.00 | 10.00 | 13.00 | 18.00 |
| ***SURPS Impulsivity[6]*** | **BL** | 145 | 0.68% | 9.77 | 1.98 | 5.00 | 8.00 | 10.00 | 11.00 | 14.00 |
| | **FU12** | 101 | 30.82% | 9.79 | 2.26 | 5.00 | 8.00 | 10.00 | 12.00 | 16.00 |
| | **FU24** | 105 | 28.08% | 9.66 | 2.18 | 5.00 | 8.00 | 10.00 | 11.00 | 15.00 |
| | **FU36** | 99 | 32.19% | 9.20 | 2.27 | 5.00 | 8.00 | 9.00 | 11.00 | 16.00 |
| ***SURPS Sensation seeking[6]*** | **BL** | 145 | 0.68% | 16.46 | 3.21 | 7.00 | 15.00 | 16.00 | 19.00 | 23.00 |
| | **FU12** | 101 | 30.82% | 16.89 | 3.24 | 10.00 | 15.00 | 17.00 | 19.00 | 23.00 |
| | **FU24** | 105 | 28.08% | 17.31 | 2.93 | 9.00 | 16.00 | 17.00 | 20.00 | 23.00 |
| | **FU36** | 99 | 32.19% | 17.34 | 3.08 | 10.00 | 15.00 | 18.00 | 19.00 | 24.00 |
| ***FTND Sum[7]*** | **BL** | 146 | 0.00% | 0.21 | 0.95 | 0.00 | 0.00 | 0.00 | 0.00 | 8.00 |
| | **FU12** | 101 | 30.82% | 0.17 | 0.60 | 0.00 | 0.00 | 0.00 | 0.00 | 4.00 |
| | **FU24** | 105 | 28.08% | 0.28 | 0.88 | 0.00 | 0.00 | 0.00 | 0.00 | 5.00 |
| | **FU36** | 44 | 69.86% | 0.75 | 1.45 | 0.00 | 0.00 | 0.00 | 1.00 | 7.00 |

* Time points include baseline (BL), follow-up six months (FU06), follow-up 12 months (FU12), etc.

[1] Total score of obsessive compulsive drinking scale (OCDS; Anton et al., 1995);

[2] Total score of alcohol dependence scale (ADS; Skinner & Allen, 1982);

[3] Barrat Impulsiveness Scale -11 (BIS; Patton et al., 1995), which assesses three subtraits impulsivity: attention, motor, and non-planning; the sum score of the three was also calculated;

[4] Drinking motive questionnaire (DMQ; Cooper, 1994), including conformity, coping, enhancement and social as four motivations for alcohol use;

[5] Brief Alcohol Expectancy Questionnaire (AEQ; Brown et al., 1987), which assesses the expected reinforcing effect of alcohol;

[6] Substance Use Risk Profile Scale (SURPS; Woicik et al., 2009), which measures substance uses risk based on four personality dimensions: anxiety sensitivity, hopelessness, impulsivity, and sensation seeking;

[7] Sum score of the Fagerström Nicotine Dependence Scale (Heatherton et al., 1991).

## A.5 fMRI Data Acquisition and Preprocessing

The imaging data was required with a Siemens 3-Tesla MRI scanner (Magnetom Trio, Siemens, Erlangen, Germany). Echo-planar imaging (EPI) sequence (TR = 2410 ms; TE = 25 ms; flip angle = 80°; voxel size = 3.0×3.0×2.0 mm with 1 mm gap; FOV = 192×192 mm; in-plane resolution: 64 × 64 pixels) consisting of 42 transversal slices was acquired in descending order with a rotation of -25° to the anterior commissure-posterior commissure line. The structural T1 weighted (Magnetization-Prepared Rapid Gradient-Echo; MPRAGE) image was also acquired (TR = 1900 ms; TE = 2.52 s; flip angle = 9°; voxel size = 1.0×1.0×1.0 mm; FOV = 256×256 mm).

Preprocessing of the fMRI data was performed with Nipype (Gorgolewski et al., 2011). The EPI images were slice time corrected, and then realigned to the first image of each time series for motion correction. Voxel displacement map was estimated based on the field maps to correct for the spatial distortion of the EPI images, after which the mean EPI images were coregistrated to the individual structural image. The individual structural image was segmented and normalized to the MNI space, and the normalization parameters were applied to the EPI images, which were resampled into a voxel size of 2×2×2 mm, and smoothed with a Gaussian Kernel with full width at half maximum of 8 mm. The high-pass filter with the width of 128 s was applied for the first-level fMRI analysis (same as Nebe et al. [2018]).

## A.6 fMRI First-Level Model & ROI Definition

These two RPE parametric modulators were our regressors of interest. The MB and MF RPE parametric modulators were calculated from the computational model, which is the same as in Daw et al. (2011) and has also been described in detail in SM1.1 in Nebe et al. (2018). In the computational model, the parameter ω that ranges from 0 to 1 represents the balance between the MF and MB control. Absolute MF control yields $\omega = 0$, while absolute MB control yields $\omega = 1$. The MF RPE signal was thus derived from the assumption of pure MF control ($\omega = 0$). The MB parametric regressor was modeled as the difference between the MB and MF RPE so that it captured only the part of the RPE that was not accounted for by the MF RPE. As regressors of no interest, an onset regressor at second-stage outcome presentation and an onset regressor at the first stage presentation were also included. The normalized measure of first stage action value and its partial derivative with respect to $\omega$ (consistent with Daw et al., 2011) were modeled as two parametric modulators for the first-stage presentation

(calculated based on the computational model). The six nuisance (motion) regressors were also included in the first-level model.

The same ROI masks were used as in Nebe et al. (Nebe et al., 2018). The vmPFC mask (Figure S2; left) was extracted from the Neurosynth database of meta-analysis (https://www.neurosynth.org/) by searching the term "vmPFC"; the anterior cingulate cortex was additionally removed from the mask. The VS mask (Figure S2; right) was extracted from the BrainMap database (Nielsen & Hansen, 2002) by searching the term "accumbens". Both masks were smoothed and binarized. The VS mask was subtracted from the vmPFC mask in order to make sure that they do not overlap with each other.

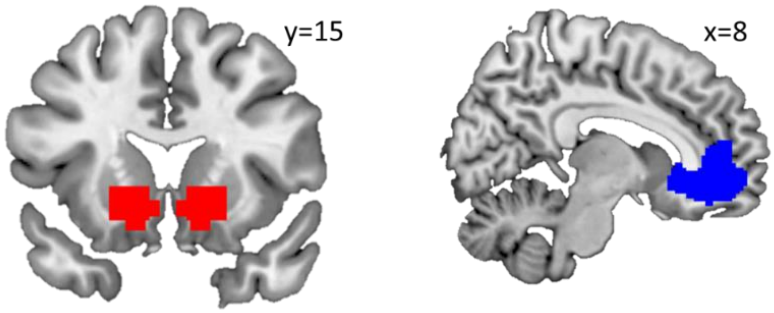


**Figure S2: Regions of interest masks**. VS (left; in red) and vmPFC (right; in blue).

## A.7 Quadratic Trajectory Model

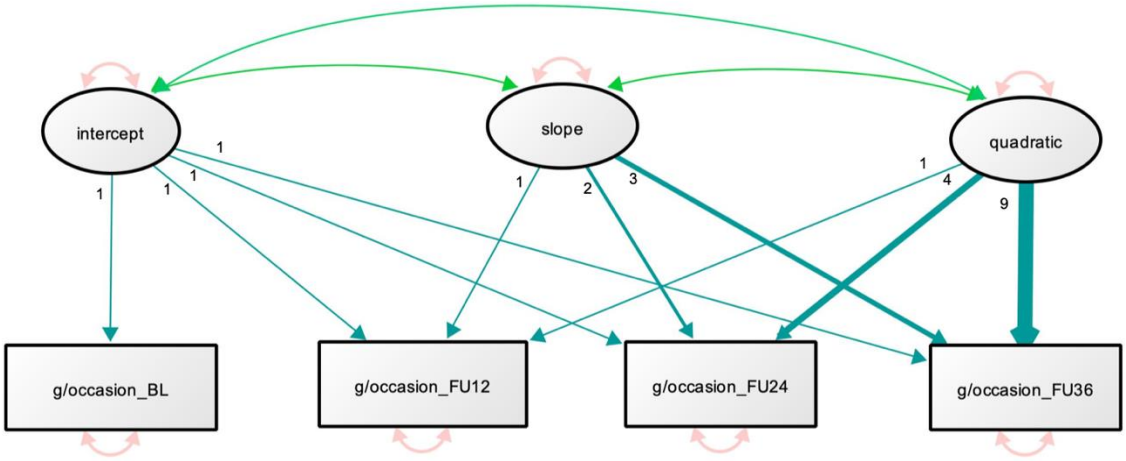


**Figure S3: Quadratic latent growth curve model.** The quadratic model for the gram/occasion variable was constructed by adding one extra quadratic latent term. The loading from the intercept, slope, and quadratic terms to the drinking variables were fixed as seen in the figure. The covariances between the intercept, slope, and quadratic terms were freely estimated. The AUDIT-C quadratic model was constructed in the same way (not shown in the graph due to redundancy). We found that adding the quadratic term did not improve model fit of either model (AUDIT-C model: $\Delta\chi2(3, 133) = -3.654$, $p = .301$; gram/occasion model: $\Delta\chi2(3, 146) = -3.271$, $p = .352$).

## A.8 Correlation Between All Two-Step Predictors

*Table S4: Correlation matrix between all the two-step predictors*

**Pearson correlations between the two-step predictors (including omega from the computational model)**

| | | Omega | MF score | MB score | MB VS | MB vmPFC | MF VS | MF vmPFC |
|---|---|---|---|---|---|---|---|---|
| **Omega** | Pearson's *r* | — | | | | | | |
| | *p* value | — | | | | | | |
| **MF score** | Pearson's *r* | 0.078 | — | | | | | |
| | *p* value | .346 | — | | | | | |
| **MB score** | Pearson's *r* | **0.654** | **-0.317** | — | | | | |
| | *p* value | **< .001** | **< .001** | — | | | | |
| **MB VS** | Pearson's *r* | 0.123 | -0.001 | 0.044 | — | | | |
| | *p* value | .14 | .99 | .595 | — | | | |
| **MB vmPFC** | Pearson's *r* | **0.255** | 0.027 | **0.215** | **0.749** | — | | |
| | *p* value | **.002** | .743 | **.009** | **< .001** | — | | |
| **MF VS** | Pearson's *r* | **0.244** | 0.017 | 0.11 | **0.272** | **0.293** | — | |
| | *p* value | **.003** | .842 | .186 | **< .001** | **< .001** | — | |
| **MF vmPFC** | Pearson's *r* | **0.187** | 0.081 | 0.159 | **0.184** | **0.231** | **0.683** | — |
| | *p* value | **.024** | .332 | .056 | **.026** | **.005** | **< .001** | — |

*Note: The significant correlations are marked in bold.*

## A.9 Controlling for Executive Functions

As seen in Table S5, the path estimates and the model fit stayed almost the same when including different executive functions. The model fit was slightly better when including the working memory variable, and the path estimation as well as its significance level changed slightly. However, the working memory variable itself did not predict the trajectory. Therefore, the working memory may modulate the association between the MB behavioural score and the binge drinking development, but only to a very limited extent.

*Table S5: Gram/occasion model estimates after including cognitive function variables*

| | | Path | Estimate | SE | *Z* | *p* value |
|---|---|---|---|---|---|---|
| **Binge drinking score (gram alcohol / drinking occasion) past year—included TMT-B** | | | | | | |
| Intercept | MF | Behavioral score | -28.018 | 19.221 | -1.458 | .145 |
| | | VS signal | -22.755 | 14.087 | -1.615 | .106 |
| | | vmPFC signal | 12.982 | 11.400 | 1.139 | .255 |
| | MB | Behavioral score | 9.604 | 12.824 | 0.749 | .454 |
| | | VS signal | -4.723 | 6.962 | -0.678 | .497 |
| | | vmPFC signal | 2.406 | 5.370 | 0.448 | .654 |
| | Processing speed | TMT-B | 0.280 | 0.194 | 1.442 | .149 |
| Slope | MF | Behavioral score | -1.696 | 6.742 | -0.252 | .801 |
| | | VS signal | 3.466 | 4.901 | 0.707 | .479 |
| | | vmPFC signal | 0.186 | 4.003 | 0.046 | .963 |
| | MB | Behavioral score | -12.971 | 4.461 | -2.908 | **.004** *** |
| | | VS signal | -0.721 | 2.451 | -0.294 | .769 |
| | | vmPFC signal | 0.640 | 1.860 | 0.344 | .731 |
| | Processing speed | TMT-B | -0.074 | 0.063 | -1.177 | .239 |
| | | model fit: $\chi 2 = 65.21$, df = 35, $p = .001$, CFI = 0.918, RMSEA = 0.077, SRMR = 0.082 | | | | |

| | | Path | Estimate | SE | *Z* | *p* value |
|---|---|---|---|---|---|---|
| **Binge drinking score (gram alcohol / drinking occasion) past year—included DSST** | | | | | | |
| Intercept | MF | Behavioral score | -28.364 | 19.211 | -1.476 | .140 |
| | | VS signal | -25.199 | 14.207 | -1.774 | .076 |
| | | vmPFC signal | 13.823 | 11.413 | 1.211 | .226 |
| | MB | Behavioral score | 10.028 | 12.865 | 0.779 | .436 |
| | | VS signal | -5.160 | 6.918 | -0.746 | .456 |
| | | vmPFC signal | 3.349 | 5.262 | 0.636 | .524 |
| | Processing speed | DSST | -0.416 | 0.279 | -1.490 | .136 |
| Slope | MF | Behavioral score | -1.473 | 6.707 | -0.220 | .826 |
| | | VS signal | 4.021 | 4.930 | 0.816 | .415 |
| | | vmPFC signal | -0.053 | 4.002 | -0.013 | .989 |
| | MB | Behavioral score | -12.978 | 4.498 | -2.886 | **.004 *** ** |
| | | VS signal | -0.569 | 2.433 | -0.234 | .815 |
| | | vmPFC signal | 0.371 | 1.825 | 0.203 | .839 |
| | Processing speed | DSST | 0.101 | 0.093 | 1.089 | .276 |
| | | model fit: $\chi2 = 66.37$, df = 35, $p = .001$, CFI = 0.915, RMSEA = 0.078, SRMR = 0.084 | | | | |
| **Binge drinking score (gram alcohol / drinking occasion) past year—included DSbw** | | | | | | |
| Intercept | MF | Behavioral score | -29.514 | 19.337 | -1.526 | .127 |
| | | VS signal | -23.187 | 14.134 | -1.641 | .101 |
| | | vmPFC signal | 13.062 | 11.487 | 1.137 | .255 |
| | MB | Behavioral score | 4.831 | 12.636 | 0.382 | .702 |
| | | VS signal | -6.309 | 6.931 | -0.910 | .363 |
| | | vmPFC signal | 4.113 | 5.282 | 0.779 | .436 |
| | Working memory | DSbw | -0.248 | 3.407 | -0.073 | .942 |
| Slope | MF | Behavioral score | -0.928 | 6.703 | -0.138 | .890 |
| | | VS signal | 2.738 | 4.923 | 0.556 | .578 |
| | | vmPFC signal | 0.057 | 3.997 | 0.014 | .989 |
| | MB | Behavioral score | -10.820 | 4.373 | -2.474 | **.013 *** |
| | | VS signal | -0.052 | 2.431 | -0.022 | .983 |
| | | vmPFC signal | 0.008 | 1.825 | 0.005 | .996 |
| | Working memory | DSbw | -1.141 | 1.152 | -0.991 | .322 |
| | | model fit: $\chi2 = 56.38$, df = 35, $p = .012$, CFI = 0.940, RMSEA = 0.065, SRMR = 0.080 | | | | |

| | | Path | Estimate | SE | *Z* | *p* value |
|---|---|---|---|---|---|---|
| **Binge drinking score (gram alcohol / drinking occasion) past year—included MWT** | | | | | | |
| Intercept | MF | Behavioral score | -28.689 | 19.137 | -1.499 | .134 |
| | | VS signal | -20.003 | 14.099 | -1.419 | .156 |
| | | vmPFC signal | 12.256 | 11.367 | 1.078 | .281 |
| | MB | Behavioral score | 6.845 | 12.387 | 0.553 | .581 |
| | | VS signal | -7.443 | 6.884 | -1.081 | .280 |
| | | vmPFC signal | 4.829 | 5.238 | 0.922 | .357 |
| | Crystallized intelligence | MWT | -1.688 | 0.931 | -1.814 | .070 |
| Slope | MF | Behavioral score | -1.338 | 6.737 | -0.199 | .843 |
| | | VS signal | 3.147 | 4.933 | 0.638 | .524 |
| | | vmPFC signal | 0.237 | 4.023 | 0.059 | .953 |
| | MB | Behavioral score | -11.770 | 4.347 | -2.707 | **.007** *** |
| | | VS signal | -0.163 | 2.441 | -0.067 | .947 |
| | | vmPFC signal | 0.153 | 1.833 | 0.083 | .934 |
| | Crystallized intelligence | MWT | -0.084 | 0.243 | -0.344 | .731 |
| | | model fit: $\chi 2 = 56.06$, df = 35, $p = 0.013$, CFI = 0.942, RMSEA = 0.064, SRMR = 0.080 | | | | |

** P value < .05 ** P value < .01*

*Color-coding: green highlights the negative association between model-based (MB) score and the slope after controlling for cognitive functions; yellow highlights the non-significant associations between the cognitive functions and the intercept as well as the slope.*

## A.10 Controlling for Impulsivity Level

In order to control for the effect of impulsivity on the drinking trajectory, we ran two additional models with respect to the two drinking variables by including the baseline BIS score measurement as a covariate along with other two-step predictors. As shown in Table S6, no parameter estimates showed much change compared to the original model, suggesting that the impulsivity level did not play a significant role in addition to the two-step predictors in predisposing the drinking trajectory development in our sample.

*Table S6: LGCM results after including baseline BIS score as a covariate*

| | | Path | Estimate | SE | *Z* | *p* value |
|---|---|---|---|---|---|---|
| **AUDIT consumption score** | | | | | | |
| Intercept | MF | Behavioral score | -1.466 | 0.961 | -1.527 | .127 |
| | | VS signal | -1.268 | 0.670 | -1.892 | .058 |
| | | vmPFC signal | 1.415 | 0.536 | 2.641 | **.008 **** |
| | MB | Behavioral score | 0.804 | 0.600 | 1.340 | .180 |
| | | VS signal | -0.309 | 0.322 | -0.960 | .337 |
| | | vmPFC signal | -0.005 | 0.244 | -0.021 | .983 |
| | Impulsivity | BIS score | 0.009 | 0.032 | 0.272 | .786 |
| Slope | MF | Behavioral score | 0.325 | 0.172 | 1.896 | .058 |
| | | VS signal | 0.260 | 0.114 | 2.290 | **.022 *** |
| | | vmPFC signal | -0.148 | 0.093 | -1.592 | .111 |
| | MB | Behavioral score | 0.129 | 0.104 | 1.240 | .215 |
| | | VS signal | -0.035 | 0.057 | -0.607 | .544 |
| | | vmPFC signal | 0.017 | 0.043 | 0.398 | .690 |
| | Impulsivity | BIS score | -0.001 | 0.006 | -0.205 | .838 |

| | | Path | Estimate | SE | *Z* | *p* value |
|---|---|---|---|---|---|---|
| **Binge drinking score (gram alcohol / drinking occasion) past year** | | | | | | |
| Intercept | MF | Behavioral score | -29.148 | 19.246 | -1.514 | .130 |
| | | VS signal | -23.187 | 14.134 | -1.641 | .101 |
| | | vmPFC signal | 12.106 | 11.443 | 1.058 | .290 |
| | MB | Behavioral score | 4.202 | 12.409 | 0.339 | .735 |
| | | VS signal | -6.408 | 6.894 | -0.930 | .353 |
| | | vmPFC signal | 4.503 | 5.263 | 0.856 | .392 |
| | Impulsivity | BIS score | 0.823 | 0.706 | 1.165 | .244 |
| Slope | MF | Behavioral score | -1.007 | 6.771 | -0.149 | .882 |
| | | VS signal | 3.337 | 4.947 | 0.675 | .500 |
| | | vmPFC signal | 0.342 | 4.030 | 0.085 | .932 |
| | MB | Behavioral score | -11.531 | 4.340 | -2.657 | **.008 **** |
| | | VS signal | -0.320 | 2.442 | -0.131 | .896 |
| | | vmPFC signal | 0.192 | 1.841 | 0.105 | .917 |
| | Impulsivity | BIS score | -0.084 | 0.243 | -0.344 | .731 |

** P value < 0.05 ** P value < 0.01*

*AUDIT consumption score model fit: $\chi^2$ = 89.12, df = 58, p = .005, CFI = 0.958, RMSEA = 0.064, SRMR = 0.075; Alcohol intake per drinking occasion past year model fit: $\chi^2$= 55.69, df = 35, p = 0.015, CFI = 0.942, RMSEA = 0.064, SRMR = 0.080*

*Color-coding: green highlights that the original significant associations stays after controlling for the BIS score; yellow highlights the non-significant associations between the BIS score and the intercept as well as the slope of the two trajectories.*

## A.11 Alcohol Expectancies Interacting With MB and MF Control in Predicting the Drinking Trajectory

In Sebold et al. 2017 (Sebold et al., 2017), it was found that the interaction between alcohol expectancies and the MB control assessed by the two-step task can predict low treatment outcome for AUD patients. Therefore, it is of interest to explore whether such an interaction already exists in our sample in predicting the risky drinking trajectory. The alcohol expectancies were assessed by the alcohol expectancy questionnaire (AEQ) (Brown et al., 1987). The brief German version consists of 19 questions that described the expected reinforcing effect of alcohol. Examples of the questions are “When I have a drink, I find it easier to open up and express my feelings” and “Alcohol reduces muscle tension”. Participants needed to indicate whether or not they agree with the statement, which was coded as either one or zero, respectively. The descriptive statistics of the AEQ score is displayed in Table S3. Following Sebold et al. 2017, we split the participants into high- ($N$ = 83) and low-AEQ ($N$ = 60) score group; three people were removed from the analysis because no valid AEQ data were available. In order to investigate how the AEQ score interacts with the MB control in predicting the drinking trajectory, we ran the gram/occasion model for the two high- and low-AEQ score groups separately. As shown in Table S7, the negative association between MB behavioral control and the slope of the gram/occasion trajectory was only seen in the high- but not the low-AEQ score group. This finding suggests that individuals with low MB control, combined high expectations of the positive reinforcing effect of alcohol, may particularly be susceptible to the development of risky binge drinking trajectory. We did not explore this association for the AUDIT-C trajectory model since we did not find an association between the MB control and the AUDIT-C trajectory; also we did not expect MF control to interact with the AEQ scores in predicting the drinking trajectory.

*Table S7: LGCM results separated for high- and low-AEQ score group*

| | | Path | Estimate | SE | *Z* | *p* value |
|---|---|---|---|---|---|---|
| **Binge drinking score (gram alcohol / drinking occasion) past year: high-AEQ group (*N* = 83)** | | | | | | |
| Intercept | MF | Behavioral score | -31.969 | 25.978 | -1.231 | .218 |
| | | VS signal | -32.051 | 20.279 | -1.581 | .114 |
| | | vmPFC signal | 6.332 | 17.125 | 0.370 | .712 |
| | MB | Behavioral score | -0.093 | 17.233 | -0.005 | .996 |
| | | VS signal | -16.354 | 11.347 | -1.441 | .150 |
| | | vmPFC signal | 8.095 | 8.134 | 0.995 | .320 |
| Slope | MF | Behavioral score | -3.865 | 8.028 | -0.481 | .630 |
| | | VS signal | 4.205 | 6.577 | 0.639 | .523 |
| | | vmPFC signal | 3.379 | 5.591 | 0.604 | .546 |
| | MB | Behavioral score | -13.489 | 5.446 | -2.477 | **.013 *** |
| | | VS signal | 2.201 | 3.703 | 0.594 | .552 |
| | | vmPFC signal | -0.609 | 2.619 | -0.233 | .816 |
| model fit: $\chi2$ = 54.78, df = 27, *p* = .001, CFI = 0.893, RMSEA = 0.111, SRMR = 0.113 | | | | | | |
| **Binge drinking score (gram alcohol / drinking occasion) past year: low-AEQ group (*N* = 60)** | | | | | | |
| Intercept | MF | Behavioral score | 7.964 | 23.982 | 0.332 | .740 |
| | | VS signal | -12.069 | 15.808 | -0.763 | .445 |
| | | vmPFC signal | 13.758 | 11.407 | 1.206 | .228 |
| | MB | Behavioral score | -0.596 | 14.047 | -0.042 | .966 |
| | | VS signal | 7.903 | 6.528 | 1.211 | .226 |
| | | vmPFC signal | -4.871 | 5.421 | -0.899 | .369 |
| Slope | MF | Behavioral score | -0.752 | 12.066 | -0.062 | .950 |
| | | VS signal | 1.710 | 7.631 | 0.224 | .823 |
| | | vmPFC signal | -0.014 | 5.754 | -0.002 | .998 |
| | MB | Behavioral score | -4.838 | 6.900 | -0.701 | .483 |
| | | VS signal | -4.210 | 3.384 | -1.244 | .214 |
| | | vmPFC signal | 2.102 | 2.698 | 0.779 | .436 |
| model fit: $\chi2$ = 51.99, df = 27, *p* = .003, CFI = 0.809, RMSEA = 0.124, SRMR = 0.139 | | | | | | |

** p value < .05*

# B Supplementary Materials: Study 2

## B.1 Error Rate Across All Experimental Conditions for High- and Low-Risk Drinkers

Before collapsing all fourteen experimental conditions (seven Pavlovian conditioned stimuli [CS] × two instrumental stimuli) into incongruent and congruent conditions, we looked at the ER across all conditions for high- and low-risk drinkers separately, which is what was also done in Sommer et al. (2020); Sommer et al. (2017). Within the trials with background associated with monetary reward or loss, ER showed a symmetric pattern between collect CS+ and leave CS-, as well as between collect CS- and leave CS+ (see Figure S4A). ER did not differ with respect to the Pavlovian cue salience (magnitude of €1 or €2). Therefore, to make the analysis more parsimonious, we merged the ten experimental conditions into congruent (i.e., CS+ collect and CS- leave) and incongruent conditions. Additionally, within the incongruent condition, the ER differences between the high- and low-risk drinkers did not differ under the "collecting a good shell with a negative background" and the "leaving a bad shell with a positive background" conditions.

In addition to the trials with backgrounds associated with monetary reward or loss, there were also trials with images of alcoholic beverages or water presented in the background. In the alcohol/water trials, the two types of instrumental shells were shown with tiled alcohol or water pictures (two types each) in the background; the sound of pouring alcohol or water into a glass was played spontaneously. There were 72 trials in total with 9 trials for each combination. Participants had a response window of 2.4 s within which they needed to give responses according to the "good "or "bad" quality of the shell. The alcohol/water trials were also performed under extinction.

We analysed the alcohol/water trials together with the neutral trials (€0) in order to see whether the alcohol or water backgrounds would elicit a different response pattern compared to the neutral trials. Mixed-effects logistic regression was performed where the responses (correct/incorrect) were regressed on the instrumental condition (collect/leave), Pavlovian background type (alcohol/water/neutral), and risk group (low-risk drinkers/high-risk drinkers). Random effects included within-subject factors (intercept, instrumental behavior, and Pavlovian background type). The logistic regression showed that the low-risk drinkers tended to give more correct responses than the high-risk drinkers across the three

conditions, but this effect was only marginally significant ($\beta$ = 0.085, $p$ = .069). No difference was seen between the instrumental conditions, different Pavlovian background, or the interaction between the instrumental condition and the Pavlovian background ($p$ > .510; Figure S4). These findings suggest that the valence of the alcohol or water backgrounds was not perceived differently from the valence of the neural Pavlovian stimuli background by our participants.

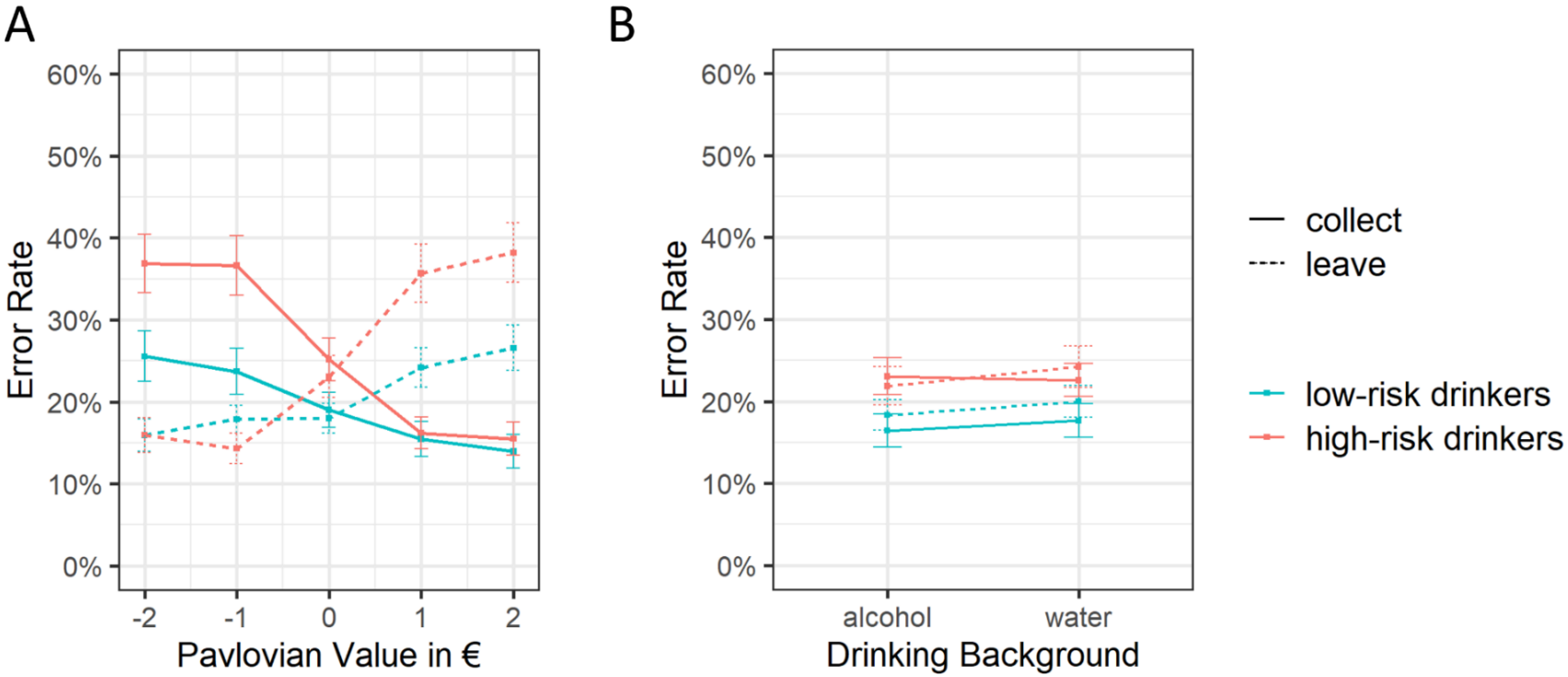


**Figure S4: Error rate across all experimental conditions for high- and low-risk drinkers.**

## B.2 Compare the Cognitive Ability Between High- and Low-Risk Drinkers

To control for the differences in the cognitive ability of the high- and low-risk drinkers, as well as its potential influence on the PIT task performance, we looked at three aspects of the cognitive ability: processing speed, working memory, and crystallized intelligence. In the Digit Symbol Substitution Task (DSST) (Wechsler, 1997), subjects needed to substitute numbers with their corresponding abstract symbols according to a list. The number of successful substitutions within 120 seconds was used as the measure of processing speed. The working memory capacity was tested by the Digit Span Backwards Test (Wechsler, 1997). The maximum number of digits a participant could recall in reverse order was used as the indicator of working memory. With respect to the crystallized intelligence, we adopted the German vocabulary test (Mehrfachwahl-Wortschatz-Intelligenztest; MWT-B) where participants needed to identify the German word among a word list including four other nonsense words (Lehrl, 2005). The individual score was the number of correct answers out of the 37-word lists. For all three measures, we compared the high- and low-risk drinking groups with two-sample *t* tests, see Table S8. No significant between-group differences were found (all $p > .49$).

*Table S8: Cognitive ability test result for high- and low-risk drinkers*

| Test | Low-risk Drinkers (*N* = 97) | | High-risk Drinkers (*N* = 94) | |
|---|---|---|---|---|
| | Mean (SD) | Range (min-max) | Mean (SD) | Range (min-max) |
| DSST | 83.98 (14.05) | 26-116 | 84.10 (11.78) | 50-120 |
| Working Memory | 8.25 (1.97) | 5-13 | 8.10 (1.97) | 5-14 |
| MWT | 24.73 (3.88) | 14-33 | 24.56 (3.82) | 14-35 |

## B.3 Behavioral PIT Effect and Generic Drinking Score

In addition to the measurement of binge drinking, we calculated a generic drinking score using M-CIDI variables, which takes into account a set of variables consisting of first-time and lifetime drinking-related variables, alcohol intake information from the last year, as well as information regarding binge drinking (Jacobi et al., 2013; Wittchen & Pfister, 1997). The calculation procedure was the same as described in the supplementary information in Nebe et al. (2018). The following seven M-CIDI variables were used for the calculation: age of first drink (in years), age of first time drunk (in years), age of first binge drinking (in years), the total number of lifetime binge drinking episodes, alcohol consumption per binge drinking episode

(gram pure ethanol/binge), alcohol consumption per occasion over the past year (gram pure ethanol/occasion) and alcohol consumption per day over the past year (g pure ethanol/day). The first three onset variables were transformed into timespan since first drink, first time being drunk, and first binge-drinking episode so that higher values indicate riskier drinking behavior. The binge-related variables were set to zero for those who did not report binge-drinking. All variables were then *z*-standardized and summed up to get the final drinking score (missing data were set to zero so that they do not influence the value of the drinking score).

Following the calculation, we ran a correlation test to examine the association between drinking score and the PIT interference effect. This dimensional drinking score showed a similar association with the PIT effect regarding the risk status variable (Figure S5): higher drinking scores were associated with higher PIT effects (Pearson's correlation $r$ = 0.12, $p_{\text{one-tailed}}$ = .048). This may suggest that the association between PIT and early hazardous alcohol use is not limited to binge drinking behavior; the association persists even when looking at more dimensional drinking behavior.

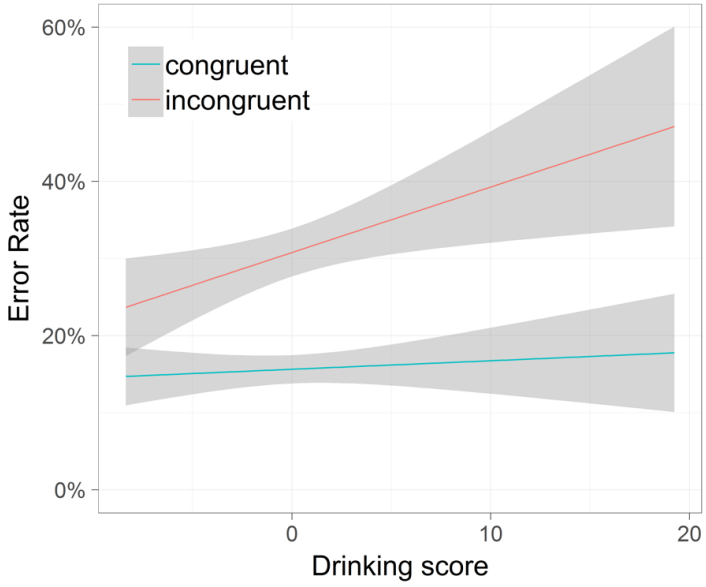


**Figure S5: The association between error rate and the drinking score showed a similar pattern to the risk status measurement.** Error rate increased more with the drinking score in the incongruent condition compared to the congruent condition ($p$ = .048).

## B.4 fMRI Data Acquisition and Preprocessing (Detailed Information)

We acquired the imaging data using a Siemens 3-Telsa MRI scanner (Magnetom Trio, Siemens, Erlangen, Germany). Echo-planar imaging (EPI) sequence (TR = 2410 ms; TE = 25 ms; flip angle = 80°; voxel size = 3.0×3.0×2.0 mm with 1 mm gap; FOV = 192×192 mm; in-plane resolution: 64×64 pixels) consisting of 42 transversal slices was acquired in descending order with a rotation of -25° to the anterior commissure-posterior commissure line. A structural T1 weighted (Magnetization-Prepared Rapid Gradient-Echo [MPRAGE]) image was also acquired (TR = 1900 ms; TE = 2.52 s; flip angle = 9°; voxel size = 1.0×1.0×1.0 mm; FOV = 256×256 mm).

Preprocessing of the fMRI data was performed with Nipype (Gorgolewski et al., 2011). The EPI images were slice time corrected and then realigned to the first image of each time series for motion correction. A voxel displacement map was estimated based on the field maps to correct for the spatial distortion of the EPI images, after which the mean EPI images were coregistered to the individual structural image. The individual structural image was segmented and normalized to the MNI space and the normalization parameters were applied to the EPI images. The EPI images were then resampled into a voxel size of 2×2×2 mm and smoothed with a Gaussian Kernel with full width at half maximum of 8 mm. The high-pass filter with the width of 128 s was applied for the first-level fMRI analysis.

## B.5 Neural Correlates of Behavioral PIT Effect – Split for High- and Low-Risk Drinkers

Following the whole-brain exploratory analysis of the neural correlates of behavioral PIT effect, we further explored whether the association between behavioral and neural incongruency effect differs with risk status. In order to test this effect statistically, we first extracted the mean parameter estimates within the activated clusters of the neural incongruency effect based on the whole-brain analysis threshold ($p < .001$, cluster size ≥ 50). After this, we performed a multiple linear regression with the mean parameter estimates within the activated clusters as the dependent variable; the risk status, behavioral PIT effect (Δ ER), as well as their interaction were used as independent variables. It is worth noting that we extracted the neural response from the activated cluster instead of the ROIs for this analysis, because we expected these clusters, as driven by the data, to give more precise neural PIT signals as compared to the ROIs. As an exploratory analysis, we ran two-tailed coefficient tests.

Linear regression showed the increase of neural activation with behavioral PIT effect was significantly stronger for the low-risk drinkers compared to high-risk drinkers in the VS ($p$ = .009) and lPFC ($p$ = .013), and this effect was marginally significant for the dmPFC ($p$ = .055), see Figure S6.

To further disentangle the neural PIT effect, we inspected the association between the neural activation in the three regions and the risk status separately for incongruent and congruent conditions. As displayed in the upper panel of Figure S7, behavioral PIT effects and the corresponding neural response were not significantly associated in the congruent condition ($p_{VS}$ = .957, $p_{lPFC}$ = .542, $p_{dmPFC}$ = .675). Both low-risk and high-risk drinkers showed the same pattern and did not significantly differ from each other (interaction effect of behavioral PIT effect and risk status: $p_{vs}$ = .744, $p_{lPFC}$ = .467, $p_{dmPFC}$ = .456). By contrast, in the incongruent condition, neural activation increased significantly with the behavioral PIT effect ($p_{VS}$ = .004, $p_{lPFC}$ = .001, $p_{dmPFC}$ = .007), and this increase was more prominent for the low-risk drinkers compared to the high-risk drinkers (interaction effect of the increase in ER and risk status: $p_{VS}$ = .012, $p_{lPFC}$ = .003, $p_{dmPFC}$ = .019). This analysis suggests that the differences between the high- and low-risk drinkers were driven by the incongruent condition instead of the congruent condition.

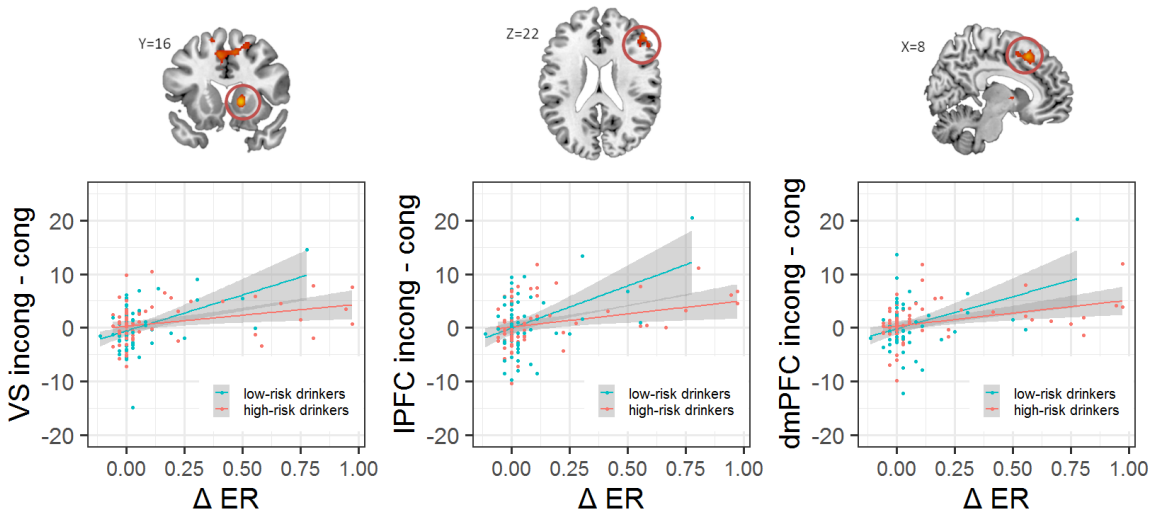


**Figure S6: Neural correlates of behavioral Pavlovian-to-instrumental transfer (PIT) effect split for risk groups.** The increase of neural activation was significantly stronger for the low-risk drinkers compared to high-risk drinkers in VS ($p$ = .009) and lPFC ($p$ = .013), and marginally significant for the dmPFC ($p$ = .055).

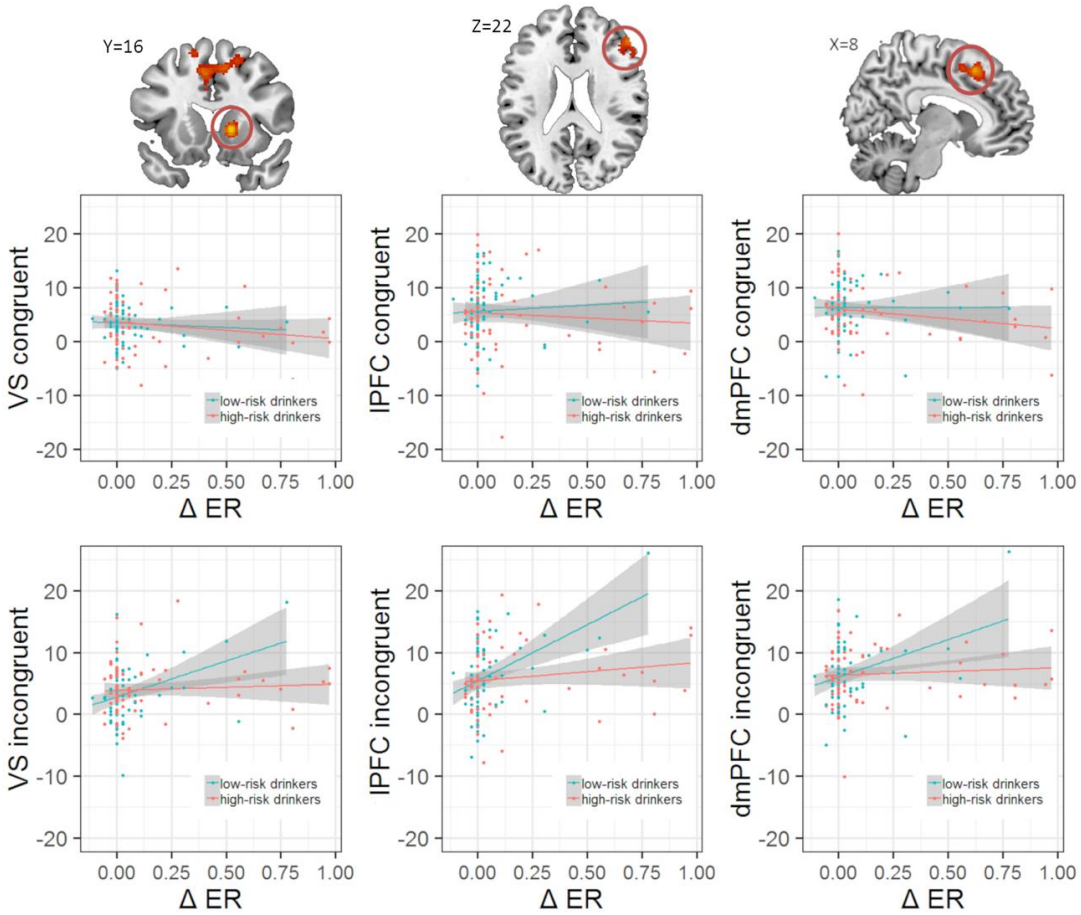


**Figure S7: Neural Pavlovian-to-instrumental transfer (PIT) effect separated for congruent and incongruent conditions, with respect to high- and low-risk drinkers.** In the congruent condition, with the PIT effect increasing, the neural activation in the VS, dmPFC, and lPFC remain stable ($p > .542$), and this effect does not differ between high- and low-risk drinkers ($p > .456$). In contrast, in the incongruent condition, behavioral PIT effect was associated with stronger neural activation in all these three regions ($p < .007$), and the association is stronger for low-risk drinkers as compared with high-risk drinkers ($p < .019$).

*Table S9: DCM Intrinsic connectivity parameter estimates for high- and low-risk drinkers*

| | Intrinsic connectivity | | | |
|---|---|---|---|---|
| | **Low-risk drinkers** | **High-risk drinkers** | **Two-sample *t* test** | |
| | *t (p)* | *t (p)* | *t* | *p* |
| VS→VS | -0.006 (.015) ** | -0.008 (.015) ** | 0.51 | .607 |
| VS→lPFC | 0.013 (.015) ** | 0.020 (.015) ** | -2.59 | .011 * |
| VS→dmPFC | 0.006 (.015) ** | 0.029 (.015) ** | -9.33 | .000 ** |
| lPFC→VS | 0.011 (.015) ** | 0.014 (.015) ** | -1.12 | .265 |
| lPFC→lPFC | -0.004 (.015) ** | -0.003 (.015) | -0.40 | .691 |
| lPFC→dmPFC | 0.017 (.015) ** | 0.013 (.015) ** | 1.74 | .084 |
| dmPFC→VS | 0.006 (.015) ** | 0.017 (.015) ** | -4.17 | .000 ** |
| dmPFC→lPFC | 0.016 (.015)** | 0.013 (.015) ** | 1.03 | .304 |
| dmPFC→dmPFC | -0.006 (.015) ** | -0.004 (.015) * | -0.77 | .443 |

* Significant at uncorrected threshold $p < .05$

** Survives Bonferroni correction for multiple comparisons (9 comparisons)

## B.6 Query Trials & Subjective Rating Analyses

During the query trials, participants responded to a series of forced-choice trials. In each trial, they were presented with two of the five monetary conditioned stimuli (CSs) that were previously paired with positive (€ 1, €2), negative (€-1, €-2) and neutral (€ 0) unconditioned stimuli (USs). Each possible pairing between the two CSs was presented three times in randomized order. Choosing CSs paired with high-value USs over the low-value USs indicated correct responses. Five participants were excluded from this analysis because no valid data for the query trials were recorded. Participants showed good performance with respect to their explicit knowledge as assessed by the accuracy (95.8% on average; 130 of 185 participants had an accuracy of 100%). The accuracy during the query trials was not associated with the behavioral interference PIT effect (error rate difference between incongruent and congruent condition; $r(184) = 0.044$; $p = .55$). The accuracy during the query trials did not differ between high- and low-risk drinkers ($T = 0.12$, df = 183, $p = .90$).

In the debriefing questionnaires after the experiment, participants needed to answer four questions for each of the five Pavlovian fractals on a scale of 1 to 7. The participants were asked to provide ratings for the following: “How pleasant you found the picture”, “How exciting you found the picture”, “How well you remembered the picture”, and “How much alcohol craving the picture triggered for you”. Nine people were excluded due to missing data,

leaving 91 subjects for both high- and low-risk drinking groups. We compared the ratings for all twenty questions between the high- and low-risk drinking groups and found no differences between the two groups with two-tailed *t* tests ($p > .063$ without correcting for multiple comparisons). The insignificant results thus suggested that the different PIT effect we saw between two groups could not be attributed to the subjective perception Pavlovian cues.

## B.7 Discussion about the Differences Between the Current Study and Garbusow, et al. 2019

Since we used the same dataset as a previously published study from our group, we would like to present a direct comparison with Garbusow et al. (2019). In this paper we investigated how the instrumental responses to the Pavlovian cue valence are coded during PIT, while we were mainly interested in how the conflict between Pavlovian cues and instrumental behavior is processed. Following the research question about the valence of Pavlovian cues, Garbusow et al. found stronger neural responses in the right Amygdala with the increased response rates to higher Pavlovian cue valence. However, the neural responses in the right Amygdala did not differ between the high- and low-risk drinkers. On the other hand, in order to investigate the conflict between Pavlovian cues and instrumental behavior, we compared the neural responses between the incongruent and congruent conditions. It was found that stronger neural responses in the ventral striatum, as well as the dorsomedial and lateral prefrontal cortices, were associated with more error rates on the behavioral level. This difference indicates that different neural mechanisms were involved when investigating the task from different perspectives. Additionally, we found differences between the high- and low-risk drinking groups in their neural responses: high-risk drinking was associated with a decreased lPFC top-down response, an increased neural response in the ventral striatum on the trend level, and weaker connectivity from the ventral striatum to the lPFC during incongruent trials. These findings provided additional insights into how the two drinking groups differ on the neural level. On the behavioral level, the two studies selected different parameters in order to optimize the different research questions addressed on the neural level. Garbusow et al. used the increase in response rates (number of button presses) with respect to the increased Pavlovian background values (from €-2 to €+2), which was calculated from the individual slopes from the general linear mixed-effect model. The response rate was chosen to capture the motivational effect of the Pavlovian cue valence. In the current study, our behavioral

outcome of interest was the error rate, which also matches our research question about conflict processing. The two behavioral outcomes are highly correlated ($r(189) = 0.96$, $p < .001$), which is not surprising since the response rate is directly associated with the error rate according to our task design—higher response rates (more button presses) indicated more collecting than leaving.

# C Supplementary Materials: Study 3

## C.1 fMRI Data Acquisition and Preprocessing

Imaging data were acquired with a Simons 3-Telsa MRI scanner (Magnetom Trio, Siemens, Erlangen, Germany). The Echo-planar imaging (EPI) sequence with TR/TE = 2410/25 ms, flip angle of 80°, voxel size = 3.0×3.0×2.0 mm with 1 mm gap, FOV = 192×192 mm, and in-plane resolution of 64×64 pixels were acquired. The EPI sequence consisted of 42 transversal slices in descending order. During the acquisition, a rotation of -25° to the anterior commissure-posterior commissure line was applied. The structural T1 weighted image was also acquired (TR/TE = 1900 ms/2.52 s; flip angle = 9°; voxel size = 1.0×1.0×1.0 mm; FOV = 256×256 mm).

We preprocessed the fMRI data with Nipype (Gorgolewski et al., 2011). The preprocessing pipeline included the following steps: slice time correction, realignment to the first image of each time series, voxel displacement to correct for the spatial distortion, coregistration to the individual structural image, segmentation and normalization to the MNI space, resampling the EPI images into a voxel size of 2×2×2 mm, smoothing with a Gaussian Kernel with full width at half maximum of 8 mm, and finally, the high-pass filter with the width of 128 s was applied.

## C.2 Descriptive Statistics for the Drinking-Related Questionnaires

*Table S10: Descriptive statistics of the drinking behaviors*

| Measures | Age | Valid | Missing Rate | Median | Mean | Std. Deviation | Minimum | Maximum |
|---|---|---|---|---|---|---|---|---|
| AUDIT consumption score (AUDIT-C) | 18.5 | 70 | 40.0% | 4.03 | 4.00 | 2.30 | 0.00 | 9.00 |
| | 19 | 80 | 32.0% | 4.13 | 4.00 | 2.07 | 0.00 | 9.00 |
| | 19.5 | 89 | 24.0% | 4.34 | 4.00 | 2.07 | 0.00 | 9.00 |
| | 20 | 90 | 23.0% | 4.44 | 4.50 | 2.19 | 0.00 | 9.00 |
| | 20.5 | 95 | 19.0% | 4.26 | 4.00 | 2.03 | 0.00 | 9.00 |
| | 21 | 117 | 0.0% | 4.27 | 4.00 | 1.97 | 0.00 | 10.00 |
| | 21.5 | 81 | 31.0% | 4.35 | 4.00 | 2.06 | 0.00 | 9.00 |
| | 22 | 83 | 29.0% | 4.39 | 4.00 | 2.18 | 0.00 | 12.00 |
| | 22.5 | 72 | 38.0% | 4.21 | 4.00 | 1.94 | 0.00 | 8.00 |
| | 23 | 70 | 40.0% | 4.16 | 4.00 | 1.98 | 0.00 | 8.00 |
| | 23.5 | 72 | 38.0% | 4.18 | 4.00 | 1.98 | 0.00 | 8.00 |
| | 24 | 70 | 40.0% | 4.24 | 4.00 | 2.18 | 0.00 | 9.00 |
| Binge drinking score (gram/drinking occasion past year) | 18 | 117 | 0.0% | 65.85 | 54 | 38.78 | 18.00 | 225.00 |
| | 19 | 102 | 13.0% | 55.09 | 45 | 36.26 | 0.00 | 180.00 |
| | 20 | 90 | 23.0% | 54.02 | 45 | 36.66 | 0.00 | 165.60 |
| | 21 | 117 | 0.0% | 42.69 | 36 | 35.18 | 0.00 | 153.00 |
| | 22 | 76 | 35.0% | 64.63 | 63.45 | 41.14 | 4.50 | 225.00 |
| | 23 | 69 | 41.0% | 57.78 | 45 | 35.94 | 0.00 | 214.20 |
| | 24 | 50 | 57.0% | 60.05 | 51.3 | 35.27 | 0.00 | 144.00 |
| Age - 1st Bingeing | - | 81 | 30.8% | 16.00 | 16.37 | 0.82 | 14.00 | 17.90 |
| Age - 1st Drink | - | 117 | 0.0% | 14.00 | 14.28 | 1.45 | 10.00 | 17.92 |
| Age - 1st Drunk | - | 113 | 3.4% | 16.00 | 15.80 | 1.10 | 12.00 | 18.01 |
| Drinking frequency past year [1] | 18 | 117 | 0.0% | 2.00 | 2.41 | 0.76 | 1.00 | 4.00 |
| | 19 | 102 | 12.8% | 3.00 | 2.57 | 0.87 | 0.00 | 4.00 |
| | 20 | 89 | 23.9% | 2.00 | 2.48 | 0.99 | 0.00 | 5.00 |
| | 21 | 117 | 0.0% | 3.00 | 2.62 | 1.04 | 1.00 | 5.00 |
| | 22 | 77 | 34.2% | 3.00 | 2.88 | 0.93 | 1.00 | 5.00 |
| | 23 | 68 | 41.9% | 3.00 | 2.85 | 0.83 | 1.00 | 5.00 |
| | 24 | 50 | 57.3% | 3.00 | 2.90 | 1.04 | 0.00 | 5.00 |
| Gram/day past year | 18 | 117 | 0.0% | 6.43 | 10.64 | 10.99 | 0.64 | 63.00 |
| | 19 | 102 | 12.8% | 5.98 | 10.70 | 10.97 | 0.00 | 66.15 |
| | 20 | 89 | 23.9% | 6.43 | 10.10 | 11.51 | 0.00 | 67.50 |
| | 21 | 117 | 0.0% | 6.43 | 10.83 | 11.01 | 0.00 | 45.00 |
| | 22 | 76 | 35.0% | 9.64 | 16.43 | 18.67 | 0.32 | 112.50 |
| | 23 | 67 | 42.7% | 9.64 | 13.53 | 11.93 | 0.19 | 45.90 |
| | 24 | 50 | 57.3% | 10.83 | 14.91 | 12.67 | 0.00 | 45.00 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Gram/bingeing occasion past year[2] | 19 | 102 | 12.8% | 103.95 | 94.73 | 77.02 | 0.00 | 427.50 |
| | 20 | 89 | 23.9% | 112.50 | 101.57 | 65.17 | 0.00 | 225.00 |
| | 22 | 76 | 35.0% | 121.50 | 113.73 | 65.91 | 0.00 | 315.00 |
| | 23 | 67 | 42.7% | 126.00 | 120.02 | 69.44 | 0.00 | 256.50 |
| | 24 | 49 | 58.1% | 135.00 | 132.89 | 86.54 | 0.00 | 427.50 |
| AUDIT total score | 18.5 | 64 | 45.3% | 5.00 | 6.22 | 4.57 | 0.00 | 25.00 |
| | 19 | 80 | 31.6% | 5.00 | 5.75 | 3.81 | 0.00 | 16.00 |
| | 19.5 | 88 | 24.8% | 5.00 | 6.34 | 4.45 | 0.00 | 25.00 |
| | 20 | 90 | 23.1% | 5.00 | 6.31 | 4.35 | 0.00 | 23.00 |
| | 20.5 | 95 | 18.8% | 5.00 | 6.02 | 4.12 | 0.00 | 22.00 |
| | 21 | 117 | 0.0% | 5.00 | 5.80 | 3.65 | 0.00 | 17.00 |
| | 21.5 | 81 | 30.8% | 5.00 | 5.72 | 3.71 | 0.00 | 19.00 |
| | 22 | 83 | 29.1% | 5.00 | 5.68 | 4.70 | 0.00 | 36.00 |
| | 22.5 | 72 | 38.5% | 4.00 | 5.18 | 3.22 | 0.00 | 16.00 |
| | 23 | 70 | 40.2% | 4.00 | 5.09 | 3.22 | 0.00 | 14.00 |
| | 23.5 | 72 | 38.5% | 5.00 | 5.31 | 3.41 | 0.00 | 18.00 |
| | 24 | 70 | 40.2% | 5.00 | 5.80 | 3.90 | 0.00 | 16.00 |

[1] Frequency assessed with six levels:
**0:** Abstinent; **1:** less than one a month; **2:** 1-3 days a month; **3:** 1-2 days a week; **4:** 3-4 days a week; **5:** (almost) daily
[2] gram/bingeing occasion during the past year were not assessed at ages 18 and 21

### C.3 fMRI Data Analysis

The first level fMRI analysis consisted of 10 onset regressors that specified the monetary PIT trials: five monetary rewards or losses (€ -2, -1, 0, 1, or 2) × two instrumental conditions (collect or leave). In addition to the monetary PIT trials, four onset regressors of no interest were also included for the alcohol/water trials (collect or leave × alcohol/water). Furthermore, all onsets button presses were included in one regressor to account for the motor responses. Finally, six motion nuisance regressors were included in the first-level model.

We defined the incongruent versus congruent contrast for each participant on the first level for our contrast of interest. Specifically, the incongruent condition consisted of the trials where positive-valenced Pavlovian cues were paired with inhibitory instrumental actions ("bad" shells with Pavlovian fractals in the background that were paired with € 1 or 2), or negative-valenced Pavlovian cues paired with approach instrumental actions ("good" shells with Pavlovian fractals that were paired with € -1 or -2). The congruent condition consisted of the trials where the Pavlovian cue valence was concordant with the instrumental action.

The first-level individual contrast images were then entered into the second-level analysis as a one-sample *t* test. As done in our baseline paper, we included the individual interference behavioral PIT effect as a covariate in the second-level model. Additionally, we included the site information (Berlin or Dresden) as a covariate of no interest to control potential site differences.

### C.4 ROI Masks

The regions of interest (ROI) were defined in the same way as our baseline report (Chen et al., 2021d). The dorsomedial prefrontal cortex (dmPFC) and lateral prefrontal cortex (lPFC) masks were generated based on a meta-analysis of cognitive inhibition (Hung et al., 2018). Based on the meta-analysis, four peaks are located in the dmPFC region (Talairach coordinates: 6/14/40; 6/26/32; 8/8/58; −6/0/54), while three peaks are located in the lPFC (Talairach coordinates: 42/26/30; 46/14/22; 52/16/14). We thus generated 12 mm spheres around each peak and used the conjunctions of the corresponding spheres for the two regions. The ventral striatum (VS) mask was defined based on a previous meta-analysis of reward-related tasks conducted with fMRI (Liu et al., 2011). We again generated two 12 mm spheres around the two peaks (MNI coordinates: −12/10/−6 and 12/10/−6) and used the conjunction as the VS mask. The three masks are displayed in Figure S8.

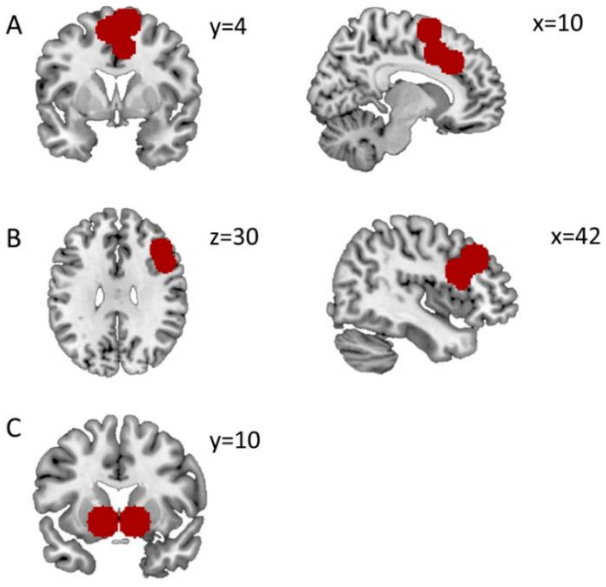


**Figure S8: Regions of interest masks**. (A) dmPFC mask (B) lPFC mask (C) VS mask.

## C.5 LGCM Model Structures

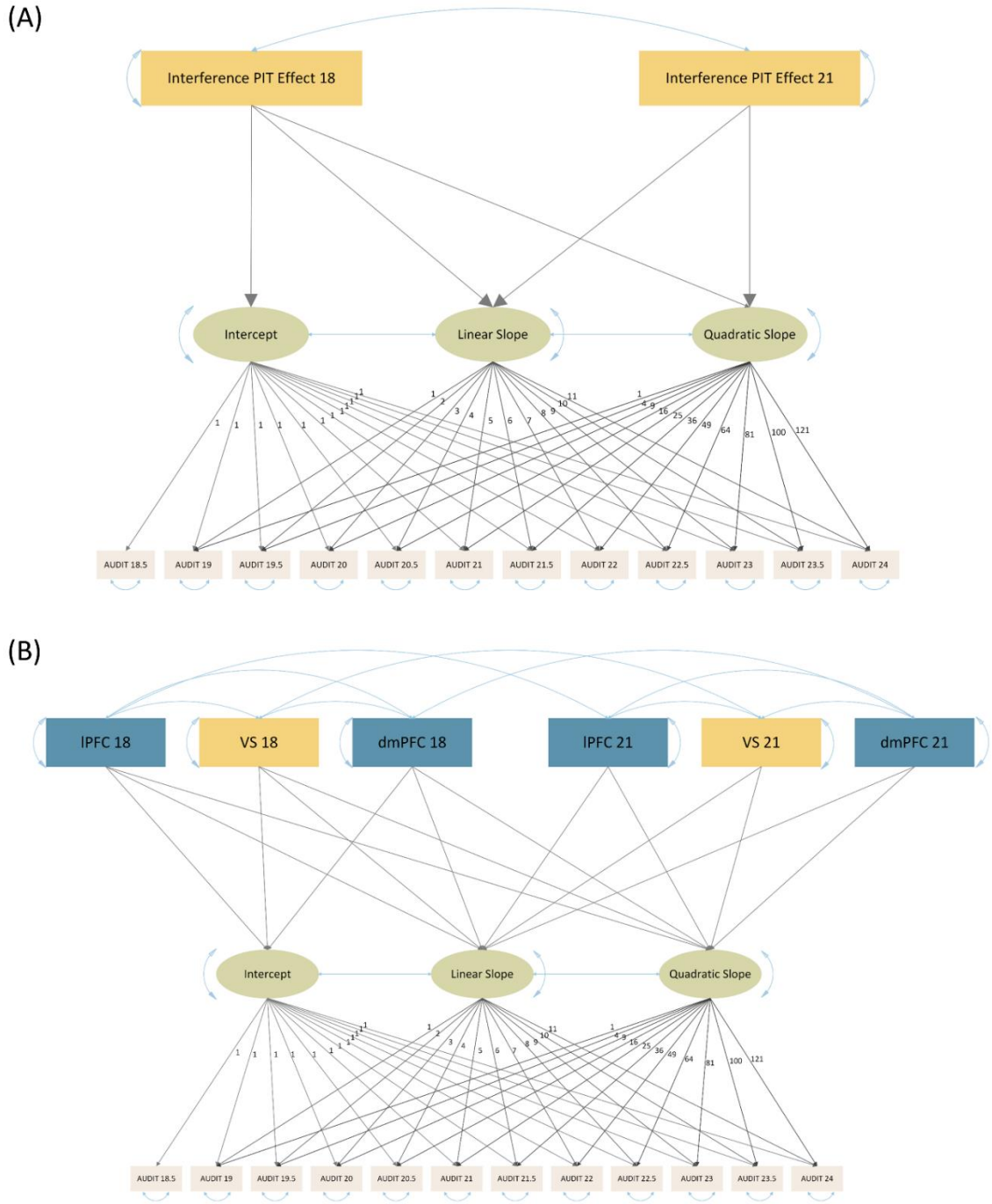


**Figure S9: Latent growth curve model structures**. (A) AUDIT-C behavioral model structure; (B) AUDIT-C neural model structure.

## C.6 Explore Clusters of AUDIT-C Developmental Trajectories

As described in the result section, the individual AUDIT-C trajectories followed developmental patterns that could be explained by the combination of the linear and quadratic slopes, and the behavioral PIT effect was associated with both slopes. Considering no changes in the AUDIT-C overtime on the group level, we suspected that distinctive clusters of participants, who followed divergent developmental trajectories in their drinking behaviors, could be identified. We could better understand the complicated associations between the behavioral PIT effect and linear and quadratic slopes by characterizing such distinctive developmental patterns. Furthermore, we could also include other questionnaire assessments in an exploratory analysis to describe the profiles of the distinctive clusters.

To achieve this, we first applied the K-means clustering method that is implemented in the machine learning module in JASP (JASP Team, 2021). The K-means clustering method is an unsupervised learning algorithm that allocates individuals into different clusters to minimize the within-cluster sum of squares. The unconditional linear and quadratic slopes (extracted from LGCM models before including any predictors) were used as input of the cluster analysis. To guarantee a sufficient power to conduct further between-cluster analyses, we set a fixed cluster size of two. After identifying the clusters, we explored differences in the behavioral PIT effect between the two clusters at ages 18 and 21 with two-sample *t* tests. We additionally calculated the individual change in the interference PIT effect from ages 18 to 21 and compared the mean difference between the two clusters. The same *t* tests were done to test whether the neural responses within the three ROIs at ages 18 and 21, as well as the change from ages 18 to 21 differ between the two clusters.

Finally, we explored whether other questionnaires of interest could characterize the cluster profiles through logistic regression. The cluster membership was included in the regression as the dependent variable. As for the predictors, in addition to the behavioral PIT effect at age 21 that was shown to be associated with the cluster membership (described in the result session), we included several drinking-related measures: family history of alcohol problems (Mann et al., 1985), the expectancies and motives of alcohol drinking as assessed by the alcohol expectancy (Brown et al., 1987) and drinking motives questionnaires, respectively (Cooper, 1994). Employment situation, educational and socioeconomic status from the sociodemographic questionnaires (Deutsche Hauptstelle für Suchtfragen, 2010), and social

readjustment rating (Holmes & Rahe, 1967) were also included as predictors. Additionally, we also tested whether the impulsivity (Meule et al., 2011; Patton et al., 1995), blatant and subtle prejudice (Pettigrew & Meertens, 1995), standardized assessment of personality (Moran et al., 2003), empathy (Davis, 1983), alexithymia (Bagby et al., 1994), and childhood trauma (Bernstein et al., 2003) were associated with the cluster membership. To avoid over-specification, for questionnaires that were available at more than one timepoint, we only included the assessments at ages 18 and 21 since these two assessments already consisted of the information about the baseline and at the turning point of the AUDIT-C trajectory. The descriptive statistics of all the included predictors at all available time points is shown in the Appendix S-5.

### C.7 Explore How Different Drinking Behaviors Are Associated with Craving and Dependence

Initially, we expected different PIT predictors to be associated with both AUDIT-C and gram/occasion trajectories in the same way. However, through the analyses, we found the associations to be different. Intriguingly, a risky drinking pattern seemed to be emerging at the later stage of young adulthood, which could be captured by the AUDIT-C but not the gram/occasion variable. We thus further investigated this difference. Considering that AUDIT-C included questions about frequency of drinking in addition to the quantity measures, as a post-hoc hypothesis, we expected the AUDIT-C to assess certain dependence and craving behaviors as individuals go further in the addiction cycle, i.e., beyond the initial binge/intoxication phase (Koob & Le Moal, 2005; Koob & Volkow, 2016). Given that both the alcohol dependence scale (ADS) (Skinner & Allen, 1982) and obsessive-compulsive drinking scale (OCDS) (Anton et al., 1995; Mann & Ackermann, 2000) were assessed every year, we calculated the correlation coefficients between AUDIT-C, gram/occasion, and the OCDS total score, as well as the ADS sum score whenever they were assessed at the same time point. These exploratory analyses were conducted to offer insights into which aspects of risky drinking behavior the AUDIT-C and gram/occasion may tap into and which aspects we could predict with the interference PIT effect.

## C.8 Descriptive Statistics of Questionnaire Measures

*Table S11: Descriptive statistics of questionnaire measures*

| Measures | Possible Range | Age | Valid | Missing Rate | Mean | Median | Std. Deviation | Minimum | Maximum |
|---|---|---|---|---|---|---|---|---|---|
| Alcohol dependence Scale (ADS) | 0-48 | 18 | 112 | 4.3% | 4.51 | 4 | 4.25 | 0 | 30 |
| | | 19 | 90 | 23.1% | 3.32 | 3 | 2.93 | 0 | 18 |
| | | 20 | 98 | 16.2% | 3.29 | 3 | 2.94 | 0 | 14 |
| | | 21 | 117 | 0.0% | 3.31 | 2 | 3.19 | 0 | 14 |
| | | 22 | 83 | 29.1% | 3.78 | 3 | 3.26 | 0 | 15 |
| | | 23 | 70 | 40.2% | 3.66 | 3 | 3.11 | 0 | 14 |
| | | 24 | 70 | 40.2% | 3.84 | 3 | 3.57 | 0 | 17 |
| Alcohol Expectancy Questionnaire (AEQ) | 19-38 | 18 | 115 | 1.7% | 28.66 | 29 | 4.51 | 19 | 38 |
| | | 19 | 88 | 24.8% | 28.88 | 30 | 4.65 | 19 | 38 |
| | | 20 | 98 | 16.2% | 28.10 | 28 | 4.68 | 19 | 38 |
| | | 21 | 117 | 0.0% | 27.86 | 28 | 4.61 | 19 | 37 |
| | | 22 | 83 | 29.1% | 28.16 | 28 | 4.24 | 19 | 37 |
| | | 23 | 70 | 40.2% | 27.80 | 27.5 | 4.97 | 19 | 38 |
| | | 24 | 70 | 40.2% | 28.33 | 27 | 4.66 | 20 | 38 |
| Barratt Impulsiveness Scale (BIS) - Short German Version | 15-60 | 18 | 116 | 0.9% | 29.72 | 30 | 4.89 | 18 | 42 |
| | | 19 | 88 | 24.8% | 28.81 | 29 | 5.59 | 18 | 44 |
| | | 20 | 98 | 16.2% | 29.07 | 29 | 5.02 | 18 | 43 |
| | | 21 | 117 | 0.0% | 29.03 | 29 | 5.10 | 17 | 45 |
| | | 22 | 83 | 29.1% | 29.22 | 29 | 5.89 | 18 | 54 |
| | | 23 | 70 | 40.2% | 27.69 | 27 | 5.87 | 16 | 42 |
| | | 24 | 69 | 41.0% | 28.38 | 28 | 5.85 | 19 | 43 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Blatant and subtle prejudice scale (BSPS) | Blatant Subscale | 10-50 | 21 | 116 | 0.9% | 15.29 | 14 | 5.34 | 10 | 42 |
| | Subtle Subscale | 10-50 | 21 | 116 | 0.9% | 30.86 | 31.5 | 7.59 | 10 | 49 |
| Childhood Trauma Questionnaire (CTQ) | Emotional Abuse | 5-25 | 21 | 117 | 0.0% | 6.08 | 5 | 1.66 | 5 | 13 |
| | Emotional Neglect | 5-25 | 21 | 117 | 0.0% | 7.72 | 7 | 2.92 | 5 | 21 |
| | Physical Abuse | 5-25 | 21 | 117 | 0.0% | 5.27 | 5 | 0.70 | 5 | 8 |
| | Physical Neglect | 5-25 | 21 | 117 | 0.0% | 5.97 | 5 | 1.67 | 5 | 13 |
| | Sexual Abuse | 5-25 | 21 | 117 | 0.0% | 5.08 | 5 | 0.46 | 5 | 9 |
| | Inconsistency Experience | 6-30 | 21 | 117 | 0.0% | 8.74 | 8 | 2.32 | 6 | 17 |
| | Trivialisation/Denial | 0-3 | 21 | 117 | 0.0% | 0.76 | 0 | 1.06 | 0 | 3 |
| Drinking Motive Questionnaire (DMQ) | Conformity Subscale | 5-25 | 18 | 115 | 1.7% | 5.88 | 5 | 1.44 | 5 | 14 |
| | | | 19 | 83 | 29.1% | 5.84 | 5 | 1.60 | 5 | 14 |
| | | | 20 | 92 | 21.4% | 6.07 | 6 | 1.44 | 5 | 11 |
| | | | 21 | 112 | 4.3% | 6.13 | 5 | 1.89 | 5 | 16 |
| | | | 22 | 80 | 31.6% | 5.85 | 5 | 1.40 | 5 | 10 |
| | | | 23 | 66 | 43.6% | 6.24 | 5 | 2.03 | 5 | 14 |
| | | | 24 | 64 | 45.3% | 6.14 | 5 | 1.88 | 5 | 12 |
| | Coping Subscale | 5-25 | 18 | 115 | 1.7% | 6.75 | 6 | 2.33 | 5 | 17 |
| | | | 19 | 83 | 29.1% | 6.12 | 6 | 1.47 | 5 | 11 |
| | | | 20 | 92 | 21.4% | 6.82 | 6 | 2.51 | 5 | 19 |
| | | | 21 | 112 | 4.3% | 6.13 | 5 | 1.50 | 5 | 10 |
| | | | 22 | 80 | 31.6% | 6.14 | 5.5 | 1.64 | 5 | 12 |
| | | | 23 | 65 | 44.4% | 6.43 | 6 | 2.02 | 5 | 16 |
| | | | 24 | 64 | 45.3% | 6.38 | 6 | 1.75 | 5 | 12 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Drinking Motive Questionnaire (DMQ) | Enhancement Subscale | 5-25 | 18 | 115 | 1.7% | 11.48 | 11 | 4.77 | 5 | 22 |
| | | | 19 | 83 | 29.1% | 10.47 | 9 | 4.40 | 5 | 23 |
| | | | 20 | 92 | 21.4% | 11.58 | 11 | 5.18 | 5 | 25 |
| | | | 21 | 112 | 4.3% | 11.18 | 10 | 4.97 | 5 | 23 |
| | | | 22 | 80 | 31.6% | 11.35 | 11 | 5.01 | 5 | 24 |
| | | | 23 | 66 | 43.6% | 11.62 | 10.5 | 5.39 | 5 | 23 |
| | | | 24 | 64 | 45.3% | 11.66 | 10.5 | 5.10 | 5 | 23 |
| | Social Subscale | 5-25 | 18 | 115 | 1.7% | 13.37 | 13 | 4.74 | 5 | 23 |
| | | | 19 | 83 | 29.1% | 11.45 | 11 | 4.43 | 5 | 23 |
| | | | 20 | 92 | 21.4% | 12.82 | 13 | 4.66 | 5 | 23 |
| | | | 21 | 112 | 4.3% | 12.81 | 12 | 4.49 | 5 | 24 |
| | | | 22 | 80 | 31.6% | 12.28 | 11 | 4.50 | 5 | 23 |
| | | | 23 | 66 | 43.6% | 12.65 | 12 | 5.01 | 5 | 25 |
| | | | 24 | 64 | 45.3% | 12.86 | 12 | 4.70 | 5 | 25 |
| Educational Status | | 0-15 | 18 | 117 | 0.0% | 1.27 | 1 | 1.19 | 1 | 13 |
| | | | 21 | 117 | 0.0% | 2.80 | 3 | 1.13 | 1 | 8 |
| Employment Situation | | 0 - Unemployed<br>1 - Student<br>2 - Pensioner<br>3 - Employed | 18 | 117 | 0.0% | 1.27 | 1 | 0.70 | 0 | 3 |
| | | | 19 | 84 | 28.2% | 1.54 | 1 | 0.99 | 0 | 3 |
| | | | 20 | 96 | 17.9% | 1.73 | 1 | 1.00 | 0 | 3 |
| | | | 21 | 117 | 0.0% | 1.74 | 1 | 1.00 | 0 | 3 |
| Family Tree Questionnaire (FTQ) | | Number of family members | 18 | 117 | 0.0% | 0.47 | 0 | 0.76 | 0 | 4 |
| | | | 21 | 114 | 2.6% | 0.90 | 1 | 1.04 | 0 | 4 |
| Obsessive Compulsive Drinking Scale (OCDS) | | 0-40 | 18 | 115 | 1.7% | 3.55 | 3 | 2.92 | 0 | 14 |
| | | | 19 | 90 | 23.1% | 2.74 | 2 | 2.57 | 0 | 11 |
| | | | 20 | 97 | 17.1% | 2.89 | 2 | 2.93 | 0 | 15 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Obsessive Compulsive Drinking Scale (OCDS) | 0-40 | 21 | 117 | 0.0% | 3.15 | 2 | 3.03 | 0 | 15 |
| | | 22 | 83 | 29.1% | 2.71 | 2 | 2.52 | 0 | 15 |
| | | 23 | 68 | 41.9% | 2.81 | 2 | 2.70 | 0 | 12 |
| | | 24 | 69 | 41.0% | 2.78 | 2 | 2.57 | 0 | 14 |
| Standardised Assessment of Personality - Abbreviated Scale (SAPAS) | 0-8 | 18 | 113 | 3.4% | 0.93 | 1 | 1.05 | 0 | 4 |
| | | 21 | 117 | 0.0% | 1.40 | 1 | 1.15 | 0 | 5 |
| Socioeconomic Status | 0 - Lower<br>1 - Lower Middle<br>2 - Middle<br>3- Upper Middle<br>4 - Upper Class | 18 | 114 | 2.6% | 2.09 | 2 | 0.65 | 0 | 4 |
| | | 19 | 80 | 31.6% | 2.09 | 2 | 0.75 | 0 | 3 |
| | | 20 | 87 | 25.6% | 2.14 | 2 | 0.69 | 0 | 4 |
| | | 21 | 106 | 9.4% | 1.98 | 2 | 0.66 | 0 | 3 |
| Empathy Scale (Saarbrücker Persönlichkeitsfragebogen zur Messung von Empathie; SPF -Empathy) | 12-60 | 21 | 117 | 0.0% | 37.10 | 37 | 6.48 | 12 | 51 |
| Social Readjustment Rating Scale (SRRS) - Holmes | 0-1466 | 18 | 117 | 0.0% | 135.89 | 114 | 95.36 | 0 | 408 |
| | | 19 | 85 | 27.4% | 185.45 | 153 | 121.57 | 0 | 514 |
| | | 20 | 97 | 17.1% | 170.49 | 139 | 125.49 | 0 | 508 |
| | | 21 | 115 | 1.7% | 167.16 | 130 | 126.38 | 0 | 561 |
| | | 22 | 83 | 29.1% | 137.17 | 101 | 105.88 | 0 | 503 |
| | | 23 | 68 | 41.9% | 140.32 | 121.5 | 106.34 | 0 | 380 |
| | | 24 | 68 | 41.9% | 164.10 | 130 | 119.70 | 0 | 556 |
| Toronto Alexithymia Scale (TAS) | 20-100 | 21 | 117 | 0.0% | 43.92 | 44 | 8.69 | 28 | 64 |

## C.9 Results of the Logistic Regression

*Table S12: Result table of the logistic regression*

| Variable (Age) | Estimate | Standard Error | Z value | P | |
|---|---|---|---|---|---|
| Intercept | 51.630 | 3487.000 | 0.015 | .988 | |
| **Interference PIT effect (21)** | **4.974** | **1.732** | **2.872** | **.004** | ** |
| Impulsivity (18) | -0.088 | 0.123 | -0.715 | .475 | |
| Impulsivity (21) | 0.029 | 0.108 | 0.268 | .789 | |
| Family Tree Questionnaire Score (18) | 0.512 | 0.706 | 0.725 | .469 | |
| Family Tree Questionnaire Score (21) | 0.623 | 0.486 | 1.282 | .200 | |
| Alcohol Expectancy (18) | -0.134 | 0.142 | -0.943 | .346 | |
| Alcohol Expectancy (21) | -0.162 | 0.116 | -1.398 | .162 | |
| Drinking Motive - Conformity (18) | -0.126 | 0.345 | -0.364 | .716 | |
| Drinking Motive - Conformity (21) | 0.220 | 0.208 | 1.057 | .291 | |
| Drinking Motive - Coping (18) | -0.087 | 0.224 | -0.386 | .699 | |
| Drinking Motive - Coping (21) | -0.500 | 0.349 | -1.432 | .152 | |
| Drinking Motive - Enhancement (18) | 0.025 | 0.159 | 0.156 | .876 | |
| Drinking Motive - Enhancement (21) | -0.110 | 0.152 | -0.724 | .469 | |
| Drinking Motive - Social (18) | -0.069 | 0.141 | -0.489 | .625 | |
| **Drinking Motive - Social (21)** | **0.379** | **0.181** | **2.091** | **.037** | * |
| Childhood Trauma - Emotional Abuse (21) | 0.068 | 0.383 | 0.177 | .860 | |
| Childhood Trauma - Physical Abuse (21) | 0.305 | 0.809 | 0.377 | .706 | |
| Childhood Trauma - Sexual Abuse (21) | -9.255 | 697.400 | -0.013 | .989 | |
| Childhood Trauma - Emotional Neglect (21) | 0.349 | 0.268 | 1.304 | .192 | |
| **Childhood Trauma - Physical Neglect (21)** | **-0.787** | **0.356** | **-2.209** | **.027** | * |
| Childhood Trauma -Trivialisation/Denial (21) | 0.608 | 0.430 | 1.414 | .157 | |
| Childhood Trauma - Inconsistency Experience (21) | 0.234 | 0.200 | 1.169 | .242 | |
| Social Readjustment (18) | 0.004 | 0.005 | 0.681 | .496 | |
| Social Readjustment (21) | -0.007 | 0.004 | -1.729 | .084 | |
| Employment Situation (18) | -0.696 | 0.833 | -0.835 | .404 | |
| Employment Situation (21) | 0.391 | 0.471 | 0.829 | .407 | |
| Education Status (18) | -0.011 | 0.808 | -0.014 | .989 | |
| Education Status (21) | -0.077 | 0.367 | -0.210 | .834 | |
| **Socioeconomic Status (18)** | **2.870** | **1.002** | **2.864** | **.004** | ** |
| Socioeconomic Status (21) | -1.654 | 0.868 | -1.905 | .057 | |
| Blatant Prejudice (21) | 0.104 | 0.118 | 0.883 | .377 | |
| Subtle Prejudice (21) | 0.018 | 0.069 | 0.263 | .792 | |
| Personality Assessment Score (18) | 0.251 | 0.405 | 0.619 | .536 | |
| Personality Assessment Score (21) | -0.383 | 0.359 | -1.065 | .287 | |
| Empathy (21) | 0.059 | 0.063 | 0.924 | .355 | |
| **Alexithymia (21)** | **-0.126** | **0.063** | **-2.005** | **.045** | * |

** $p < .01$; *$p < .05$; All significant results are displayed in bold.

## C.10 Different Types of AUDIT-C Trajectory

Figure S10 displays six types of AUDIT-C trajectories, with the intercept, linear and quadratic slopes shown in the legend. *Type 1* has a high starting point and decreases over time, as indicated in the negative linear slope. Although the quadratic slope is positive, the turning is beyond age 24; therefore, a positive trend is not visible within the six years. In contrast, *type 2* starts at a low level, and the positive linear term drives this trajectory until around age 23.5; after this, the negative quadratic term drives this trajectory down. *Type 3-6* start at the same level. *Type 3*, as compared with type 1, followed a slighter decrease over time; the turning point is also beyond age 24. *Type 4* first increases and then decreases, driven by the positive linear and negative quadratic slope. Conversely, for *Type 6*, this pattern is the opposite since it has a negative linear but positive quadratic slope. As for *Type 5*, it is an increasing function since both linear and quadratic slopes are positive.

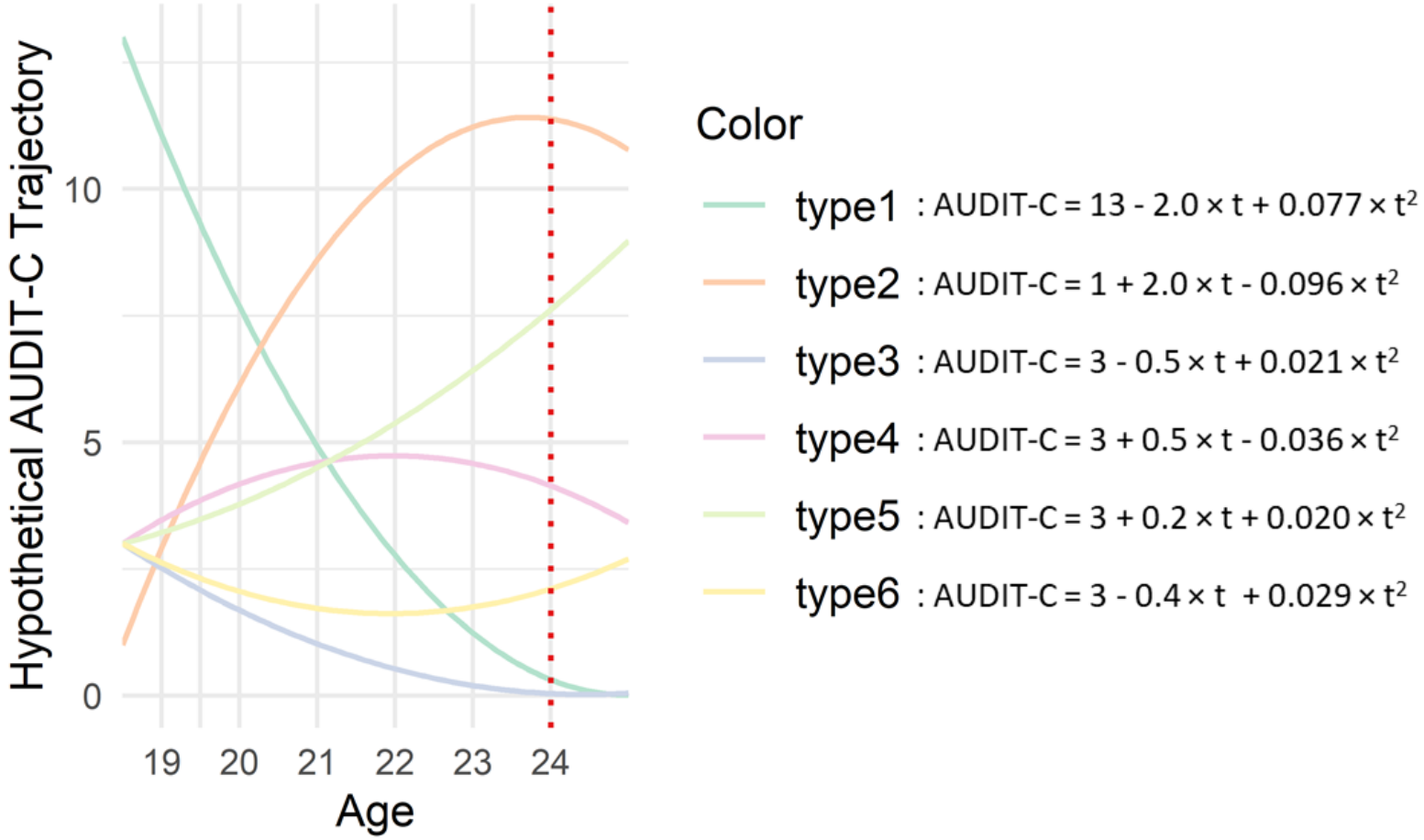


**Figure S10: Six types of Alcohol Use Disorders Identification Test consumption score (AUDIT-C) trajectories according to different combinations of the intercept, linear and quadratic slopes.**

# References


Allman, M. J., DeLeon, I. G., Cataldo, M. F., Holland, P. C., & Johnson, A. W. (2010). Learning processes affecting human decision making: An assessment of reinforcer-selective Pavlovian-to-instrumental transfer following reinforcer devaluation. *Journal of Experimental Psychology: Animal Behavior Processes*, *36*(3), 402-408. https://doi.org/10.1037/a0017876

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, *7*(4), 268-277. https://doi.org/10.1038/nrn1884

Anton, R. F., Moak, D. H., & Latham, P. (1995). The Obsessive Compulsive Drinking Scale: a self‐rated instrument for the quantification of thoughts about alcohol and drinking behavior. *Alcoholism: Clinical and Experimental Research*, *19*(1), 92-99. https://doi.org/10.1111/j.1530-0277.1995.tb01475.x

Arbuckle, J. L., Marcoulides, G. A., & Schumacker, R. E. (1996). Full information estimation in the presence of incomplete data. *Advanced Structural Equation Modeling: Issues and Techniques*, *243*, 277. https://doi.org/10.1038/npp.2009.131

Arnett, J. J. (2000). Emerging adulthood - A theory of development from the late teens through the twenties. *American Psychologist*, *55*(5), 469-480. https://doi.org/10.1037/0003-066x.55.5.469

Arnett, J. J. (2005). The developmental context of substance use in emerging adulthood. *Journal of Drug Issues*, *35*(2), 235-253. https://doi.org/10.1177/002204260503500202

Bagby, R. M., Parker, J. D., & Taylor, G. J. (1994). The twenty-item Toronto Alexithymia Scale—I. Item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research*, *38*(1), 23-32. https://doi.org/10.1016/0022-3999(94)90005-1

Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., & Andrews, T. J. (2021a). *Power Contour Estimation*. https://shiny.york.ac.uk/powercontours/

Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., & Andrews, T. J. (2021b). Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. *Psychological Methods*, *26*(3), 295-314. https://doi.org/10.1037/met0000337

Balleine, B. W., & O'Doherty, J. P. (2010). Human and rodent homologies in action control: corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, *35*(1), 48-69. https://doi.org/10.1038/npp.2009.131

Balleine, B. W., & Ostlund, S. B. (2007). Still at the choice‐point: action selection and initiation in instrumental conditioning. *Annals of the New York Academy of Sciences*, *1104*(1), 147-171. https://doi.org/10.1196/annals.1390.006

Barto, A. G., Sutton, R. S., & Watkins, C. J. C. (1989). Sequential decision problems and neural networks. *Advances in Neural Information Processing Systems*, *2*.

Bastin, J., Deman, P., David, O., Gueguen, M., Benis, D., Minotti, L., Hoffman, D., Combrisson, E., Kujala, J., & Perrone-Bertolotti, M. (2016). Direct recordings from human anterior insula reveal its leading role within the error-monitoring network. *Cerebral Cortex*, bhv352. https://doi.org/10.1093/cercor/bhv352

Beck, A., Wüstenberg, T., Genauck, A., Wrase, J., Schlagenhauf, F., Smolka, M. N., Mann, K., & Heinz, A. (2012). Effect of brain structure, brain function, and brain connectivity on relapse in alcohol-dependent patients. *Archives of General Psychiatry*, *69*(8), 842-852. https://doi.org/10.1001/archgenpsychiatry.2011.2026

Behrendt, S., Wittchen, H.-U., Höfler, M., Lieb, R., & Beesdo, K. (2009). Transitions from first substance use to substance use disorders in adolescence: is early onset associated with a rapid escalation? *Drug and Alcohol Dependence*, *99*(1-3), 68-78. https://doi.org/10.1016/j.drugalcdep.2008.06.014

Belanger, M. J., Chen, H., Hentschel, A., Garbusow, M., Ebrahimi, C., Knorr, F. G., Zech, H. G., Pilhatsch, M., Heinz, A., & Smolka, M. N. (under review). *Development of novel tasks to assess outcome-specific and general Pavlovian-to-Instrumental Transfer in humans*.

Belin, D., Belin-Rauscent, A., Everitt, B. J., & Dalley, J. W. (2016). In search of predictive endophenotypes in addiction: insights from preclinical research. *Genes, Brain, and Behaviour*, *15*(1), 74-88. https://doi.org/10.1111/gbb.12265

Belin, D., Belin-Rauscent, A., Murray, J. E., & Everitt, B. J. (2013). Addiction: failure of control over maladaptive incentive habits. *Current Opinion in Neurobiology*, *23*(4), 564-572. https://doi.org/10.1016/j.conb.2013.01.025

Bellman, R. (1957). Dynamic programming, princeton univ. *Press Princeton, New Jersey*.

Bennett, D., Niv, Y., & Langdon, A. J. (2021). Value-free reinforcement learning: policy optimization as a minimal model of operant behavior. *Current Opinion in Behavioral Sciences*, *41*, 114-121. https://doi.org/10.1016/j.cobeha.2021.04.020

Berg, N., Kiviruusu, O., Karvonen, S., Kestila, L., Lintonen, T., Rahkonen, O., & Huurre, T. (2013). A 26-year follow-up study of heavy drinking trajectories from adolescence to mid-adulthood and adult disadvantage. *Alcohol and Alcoholism*, *48*(4), 452-457. https://doi.org/10.1093/alcalc/agt026

Bernstein, D. P., Stein, J. A., Newcomb, M. D., Walker, E., Pogge, D., Ahluvalia, T., Stokes, J., Handelsman, L., Medrano, M., & Desmond, D. (2003). Development and validation of a brief screening version of the Childhood Trauma Questionnaire. *Child Abuse & Neglect*, *27*(2), 169-190. https://doi.org/10.1016/s0145-2134(02)00541-0

Berridge, K. C., & Robinson, T. E. (2016). Liking, wanting, and the incentive-sensitization theory of addiction. *American Psychologist*, *71*(8), 670-679. https://doi.org/10.1037/amp0000059

Bertsekas, D. P., & Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Athena Scientific. https://doi.org/10.1109/CDC.1995.478953

Bickel, W. K., Mellis, A. M., Snider, S. E., Athamneh, L. N., Stein, J. S., & Pope, D. A. (2018). 21st century neurobehavioral theories of decision making in addiction: Review and evaluation. *Pharmacology, Biochemistry, and Behavior*, *164*, 4-21. https://doi.org/10.1016/j.pbb.2017.09.009

Botvinick, M. M., Niv, Y., & Barto, A. G. (2009). Hierarchically organized behavior and its neural foundations: a reinforcement learning perspective. *Cognition*, *113*(3), 262-280. https://doi.org/10.1016/j.cognition.2008.08.011

Bray, S., Rangel, A., Shimojo, S., Balleine, B., & O'Doherty, J. P. (2008). The neural mechanisms underlying the influence of pavlovian cues on human decision making. *Journal of Neuroscience*, *28*(22), 5861-5866. https://doi.org/10.1523/JNEUROSCI.0897-08.2008

Brown, S. A., Christiansen, B. A., & Goldman, M. S. (1987). The Alcohol Expectancy Questionnaire: an instrument for the assessment of adolescent and adult alcohol expectancies. *Journal of Studies on Alcohol*, *48*(5), 483-491. https://doi.org/10.15288/jsa.1987.48.483

Brown, S. A., McGue, M., Maggs, J., Schulenberg, J., Hingson, R., Swartzwelder, S., Martin, C., Chung, T., Tapert, S. F., Sher, K., Winters, K. C., Lowman, C., & Murphy, S. (2008). A developmental perspective on alcohol and youths 16 to 20 years of age. *Pediatrics*, *121 Suppl 4*(Supplement 4), S290-310. https://doi.org/10.1542/peds.2007-2243D

Brown, V. M., Chen, J., Gillan, C. M., & Price, R. B. (2020). Improving the Reliability of Computational Analyses: Model-Based Planning and Its Relationship With Compulsivity. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *5*(6), 601-609. https://doi.org/10.1016/j.bpsc.2019.12.019

Brumback, T., Worley, M., Nguyen-Louie, T. T., Squeglia, L. M., Jacobus, J., & Tapert, S. F. (2016). Neural predictors of alcohol use and psychopathology symptoms in adolescents.

*Development and Psychopathology*, *28*(4), 1209-1216. https://doi.org/10.1017/S0954579416000766

Busemeyer, J. R., Gluth, S., Rieskamp, J., & Turner, B. M. (2019). Cognitive and Neural Bases of Multi-Attribute, Multi-Alternative, Value-based Decisions. *Trends in Cognitive Sciences*, *23*(3), 251-263. https://doi.org/10.1016/j.tics.2018.12.003

Cao, Z., Bennett, M., O'Halloran, L., Pragulbickaite, G., Flanagan, L., McHugh, L., & Whelan, R. (2021). Aberrant reward prediction errors in young adult at-risk alcohol users. *Addiction Biology*, *26*(1), e12873. https://doi.org/10.1111/adb.12873

Carbia, C., Lopez-Caneda, E., Corral, M., & Cadaveira, F. (2018). A systematic review of neuropsychological studies involving young binge drinkers. *Neuroscience & Biobehavioral Reviews*, *90*, 332-349. https://doi.org/10.1016/j.neubiorev.2018.04.013

Carter, B. L., & Tiffany, S. T. (1999). Meta-analysis of cue-reactivity in addiction research. *Addiction*, *94*(3), 327-340. https://www.ncbi.nlm.nih.gov/pubmed/10605857

Cartoni, E., Balleine, B., & Baldassarre, G. (2016). Appetitive Pavlovian-instrumental Transfer: A review. *Neuroscience & Biobehavioral Reviews*, *71*, 829-848. https://doi.org/10.1016/j.neubiorev.2016.09.020

Cavanagh, J. F., Eisenberg, I., Guitart-Masip, M., Huys, Q., & Frank, M. J. (2013). Frontal theta overrides pavlovian learning biases. *Journal of Neuroscience*, *33*(19), 8541-8548. https://doi.org/10.1523/JNEUROSCI.5754-12.2013

Chartier, K. G., Hesselbrock, M. N., & Hesselbrock, V. M. (2010). Development and Vulnerability Factors in Adolescent Alcohol Use. *Child and Adolescent Psychiatric Clinics of North America*, *19*(3), 493-+. https://doi.org/10.1016/j.chc.2010.03.004

Chen, C. M., Dufour, M. C., & Yi, H.-Y. (2004). Alcohol consumption among young adults ages 18–24 in the United States: Results from the 2001–2002 NESARC survey. *Alcohol Research & Health*, *28*(4), 269.

Chen, G., Padmala, S., Chen, Y., Taylor, P. A., Cox, R. W., & Pessoa, L. (2021a). To pool or not to pool: Can we ignore cross-trial variability in FMRI? *Neuroimage*, *225*, 117496. https://doi.org/10.1016/j.neuroimage.2020.117496

Chen, G., Pine, D. S., Brotman, M. A., Smith, A. R., Cox, R. W., & Haller, S. P. (2021b). Trial and error: A hierarchical modeling approach to test-retest reliability. *Neuroimage*, *245*, 118647. https://doi.org/10.1016/j.neuroimage.2021.118647

Chen, H., Mojtahedzadeh, N., Belanger, M. J., Nebe, S., Kuitunen-Paul, S., Sebold, M., Garbusow, M., Huys, Q. J. M., Heinz, A., Rapp, M. A., & Smolka, M. N. (2021c). Model-Based and Model-Free Control Predicts Alcohol Consumption Developmental Trajectory in Young Adults: A 3-Year Prospective Study. *Biological Psychiatry*, *89*(10), 980-989. https://doi.org/10.1016/j.biopsych.2021.01.009

Chen, H., Nebe, S., Mojtahedzadeh, N., Kuitunen‐Paul, S., Garbusow, M., Schad, D. J., Rapp, M. A., Huys, Q. J. M., Heinz, A., & Smolka, M. N. (2021d). Susceptibility to interference between Pavlovian and instrumental control is associated with early hazardous alcohol use. *Addiction Biology*, e12983. https://doi.org/10.1111/adb.12983

Cohen, J. Y., Haesler, S., Vong, L., Lowell, B. B., & Uchida, N. (2012). Neuron-type-specific signals for reward and punishment in the ventral tegmental area. *Nature*, *482*(7383), 85-88. https://doi.org/10.1038/nature10754

Collins, A. G. E., & Cockburn, J. (2020). Beyond dichotomies in reinforcement learning. *Nature Reviews Neuroscience*, *21*(10), 576-586. https://doi.org/10.1038/s41583-020-0355-6

Colwill, R. M., & Rescorla, R. A. (1988). Associations between the Discriminative Stimulus and the Reinforcer in Instrumental Learning. *Journal of Experimental Psychology: Animal Behavior Processes*, *14*(2), 155-164. https://doi.org/10.1037/0097-7403.14.2.155

Conner, K. R., & Bagge, C. L. (2019). Suicidal behavior: links between alcohol use disorder and acute use of alcohol. *Alcohol Research: Current Reviews*, *40*(1). https://doi.org/10.35946/arcr.v40.1.02

Cooper, M. L. (1994). Motivations for alcohol use among adolescents: Development and validation of a four-factor model. *Psychological Assessment*, *6*(2), 117. https://doi.org/10.1037/1040-3590.6.2.117

Cooper, S. R., Jackson, J. J., Barch, D. M., & Braver, T. S. (2019). Neuroimaging of individual differences: A latent variable modeling perspective. *Neuroscience & Biobehavioral Reviews*, *98*, 29-46. https://doi.org/10.1016/j.neubiorev.2018.12.022

Corbit, L. H., Janak, P. H., & Balleine, B. W. (2007). General and outcome-specific forms of Pavlovian-instrumental transfer: the effect of shifts in motivational state and inactivation of the ventral tegmental area. *European Journal of Neuroscience*, *26*(11), 3141-3149. https://doi.org/10.1111/j.1460-9568.2007.05934.x

Cservenka, A., & Brumback, T. (2017). The Burden of Binge and Heavy Drinking on the Brain: Effects on Adolescent and Young Adult Neural Structure and Function. *Frontiers in Psychology*, *8*, 1111. https://doi.org/10.3389/fpsyg.2017.01111

da Silva, C. F., & Hare, T. A. (2020). Humans primarily use model-based inference in the two-stage task. *Nature Human Behaviour*, *4*(10), 1053-1066. https://doi.org/10.1038/s41562-020-0905-y

Davis, M. H. (1983). Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, *44*(1), 113. https://doi.org/10.1037/0022-3514.44.1.113

Daw, N. D. (2018). Are we of two minds? *Nature Neuroscience*, *21*(11), 1497-1499. https://doi.org/10.1038/s41593-018-0258-2

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204-1215. https://doi.org/10.1016/j.neuron.2011.02.027

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704-1711. https://doi.org/10.1038/nn1560

Daw, N. D., & O'Doherty, J. P. (2014). Multiple systems for value learning. In *Neuroeconomics* (pp. 393-410). Elsevier.

Dawson, D. A. (2011). Defining risk drinking. *Alcohol Research & Health*, *34*(2), 144-156. https://www.ncbi.nlm.nih.gov/pubmed/22330212

Dawson, D. A., Grant, B. F., Stinson, F. S., & Zhou, Y. (2005). Effectiveness of the derived Alcohol Use Disorders Identification Test (AUDIT-C) in screening for alcohol use disorders and risk drinking in the US general population. *Alcoholism: Clinical and Experimental Research*, *29*(5), 844-854. https://doi.org/10.1097/01.alc.0000164374.32229.a2

Dayan, P., & Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: revaluation, revision, and revelation. *Cognitive, Affective, & Behavioral Neuroscience*, *14*(2), 473-492. https://doi.org/10.3758/s13415-014-0277-8

de Wit, S., Kindt, M., Knot, S. L., Verhoeven, A. A., Robbins, T. W., Gasull-Camos, J., Evans, M., Mirza, H., & Gillan, C. M. (2018). Shifting the balance between goals and habits: Five failures in experimental habit induction. *Journal of Experimental Psychology: General*, *147*(7), 1043. https://doi.org/10.1037/xge0000402

Deserno, L., Wilbertz, T., Reiter, A., Horstmann, A., Neumann, J., Villringer, A., Heinze, H. J., & Schlagenhauf, F. (2015). Lateral prefrontal model-based signatures are reduced in healthy individuals with high trait impulsivity. *Translational Psychiatry*, *5*, e659. https://doi.org/10.1038/tp.2015.139

Deutsche Hauptstelle für Suchtfragen. (2010). Deutscher Kerndatensatz zur Dokumentation im Bereich der Suchtkrankenhilfe. In *Definitionen und Erläuterungen zum Gebrauch. Stand: 05.20.2010*. IFT. https://www.suchthilfe.de/basis/kds_10_2010.pdf

Dickinson, A. (1985). Actions and habits: the development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, *308*(1135), 67-78. https://doi.org/10.1098/rstb.1985.0010

Dickinson, A., & Balleine, B. (1994). Motivational Control of Goal-Directed Action. *Animal Learning & Behavior*, *22*(1), 1-18. https://doi.org/10.3758/Bf03199951

Domenech, P., & Koechlin, E. (2015). Executive control and decision-making in the prefrontal cortex. *Current Opinion in Behavioral Sciences*, *1*, 101-106. https://doi.org/10.1016/j.cobeha.2014.10.007

Doñamayor, N., Strelchuk, D., Baek, K., Banca, P., & Voon, V. (2018). The involuntary nature of binge drinking: goal directedness and awareness of intention. *Addiction Biology*, *23*(1), 515-526. https://doi.org/10.1111/adb.12505

Dorfman, H. M., & Gershman, S. J. (2019). Controllability governs the balance between Pavlovian and instrumental action selection. *Nature Communications*, *10*(1), 1-8. https://doi.org/10.1038/s41467-019-13737-7

Du, Y., Krakauer, J. W., & Haith, A. M. (2022). The relationship between habits and motor skills in humans. *Trends in Cognitive Sciences*, *26*(5), 371-387. https://doi.org/10.1016/j.tics.2022.02.002

Dubois, J., & Adolphs, R. (2016). Building a Science of Individual Differences from fMRI. *Trends in Cognitive Sciences*, *20*(6), 425-443. https://doi.org/10.1016/j.tics.2016.03.014

Duncan, T. E., & Duncan, S. C. (2009). The ABC's of LGM: An Introductory Guide to Latent Variable Growth Curve Modeling. *Social and Personality Psychology Compass*, *3*(6), 979-991. https://doi.org/10.1111/j.1751-9004.2009.00224.x

Eder, A. B., & Dignath, D. (2016a). Asymmetrical effects of posttraining outcome revaluation on outcome-selective Pavlovian-to-instrumental transfer of control in human adults. *Learning and Motivation*, *54*, 12-21. https://doi.org/10.1016/j.lmot.2016.05.002

Eder, A. B., & Dignath, D. (2016b). Cue-elicited food seeking is eliminated with aversive outcomes following outcome devaluation. *Quarterly Journal of Experimental Psychology*, *69*(3), 574-588. https://doi.org/10.1080/17470218.2015.1062527

Egervari, G., Ciccocioppo, R., Jentsch, J. D., & Hurd, Y. L. (2018). Shaping vulnerability to addiction - the contribution of behavior, neural circuits and molecular mechanisms. *Neuroscience and Biobehavioral Reviews*, *85*, 117-125. https://doi.org/10.1016/j.neubiorev.2017.05.019

Egner, T., & Hirsch, J. (2005). Cognitive control mechanisms resolve conflict through cortical amplification of task-relevant information. *Nature Neuroscience*, *8*(12), 1784-1790. https://doi.org/10.1038/nn1594

Elliott, M. L., Knodt, A. R., & Hariri, A. R. (2021). Striving toward translation: strategies for reliable fMRI measurement. *Trends in Cognitive Sciences*, *25*(9), 776-787. https://doi.org/10.1016/j.tics.2021.05.008

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T. E., Caspi, A., & Hariri, A. R. (2020). What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. *Psychological Science*, *31*(7), 792-806. https://doi.org/10.1177/0956797620916786

Ereira, S., Pujol, M., Guitart-Masip, M., Dolan, R. J., & Kurth-Nelson, Z. (2021). Overcoming Pavlovian bias in semantic space. *Scientific Reports*, *11*(1), 3416. https://doi.org/10.1038/s41598-021-82889-8

Euston, D. R., Gruber, A. J., & McNaughton, B. L. (2012). The role of medial prefrontal cortex in memory and decision making. *Neuron*, *76*(6), 1057-1070. https://doi.org/10.1016/j.neuron.2012.12.002

Evans-Polce, R. J., Maggs, J. L., Staff, J., & Lanza, S. T. (2017). The age-varying association of student status with excessive alcohol use: ages 18 to 30 years. *Alcoholism: Clinical and Experimental Research*, *41*(2), 407-413. https://doi.org/10.1111/acer.13294

Everitt, B. J., & Robbins, T. W. (2016). Drug Addiction: Updating Actions to Habits to Compulsions Ten Years On. *Annual Review of Psychology*, *67*, 23-50. https://doi.org/10.1146/annurev-psych-122414-033457

Flagel, S. B., Clark, J. J., Robinson, T. E., Mayo, L., Czuj, A., Willuhn, I., Akers, C. A., Clinton, S. M., Phillips, P. E. M., & Akil, H. (2011). A selective role for dopamine in stimulus-reward learning. *Nature*, *469*(7328), 53-U63. https://doi.org/10.1038/nature09588

Flores-Barrera, E., Thomases, D. R., Heng, L.-J., Cass, D. K., Caballero, A., & Tseng, K. Y. (2014). Late adolescent expression of GluN2B transmission in the prefrontal cortex is input-specific and requires postsynaptic protein kinase A and D1 dopamine receptor signaling. *Biological Psychiatry*, *75*(6), 508-516. https://doi.org/10.1016/j.biopsych.2013.07.033

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E. J. (2016). Sequential Sampling Models in Cognitive Neuroscience: Advantages, Applications, and Extensions. *Annual Review of Psychology*, *67*, 641-666. https://doi.org/10.1146/annurev-psych-122414-033645

Freeman, S. M., Alvernaz, D., Tonnesen, A., Linderman, D., & Aron, A. R. (2015). Suppressing a motivationally-triggered action tendency engages a response control mechanism that prevents future provocation. *Neuropsychologia*, *68*, 218-231. https://doi.org/10.1016/j.neuropsychologia.2015.01.016

Freeman, S. M., Razhas, I., & Aron, A. R. (2014). Top-down response suppression mitigates action tendencies triggered by a motivating stimulus. *Current Biology*, *24*(2), 212-216. https://doi.org/10.1016/j.cub.2013.12.019

Friedel, E., Koch, S. P., Wendt, J., Heinz, A., Deserno, L., & Schlagenhauf, F. (2014). Devaluation and sequential decisions: linking goal-directed and model-based behavior. *Frontiers in Human Neuroscience*, *8*, 587. https://doi.org/10.3389/fnhum.2014.00587

Friston, K. J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *Neuroimage*, *19*(4), 1273-1302. https://doi.org/10.1016/s1053-8119(03)00202-7

Fröhner, J. H., Teckentrup, V., Smolka, M. N., & Kroemer, N. B. (2019). Addressing the reliability fallacy in fMRI: Similar group effects may arise from unreliable individual effects. *Neuroimage*, *195*, 174-189. https://doi.org/10.1016/j.neuroimage.2019.03.053

Gan, G., Guevara, A., Marxen, M., Neumann, M., Junger, E., Kobiella, A., Mennigen, E., Pilhatsch, M., Schwarz, D., Zimmermann, U. S., & Smolka, M. N. (2014). Alcohol-Induced Impairment of Inhibitory Control Is Linked to Attenuated Brain Responses in Right Fronto-Temporal Cortex. *Biological Psychiatry*, *76*(9), 698-707. https://doi.org/10.1016/j.biopsych.2013.12.017

Garbusow, M., Nebe, S., Sommer, C., Kuitunen-Paul, S., Sebold, M., Schad, D. J., Friedel, E., Veer, I. M., Wittchen, H. U., Rapp, M. A., Ripke, S., Walter, H., Huys, Q. J. M., Schlagenhauf, F., Smolka, M. N., & Heinz, A. (2019). Pavlovian-To-Instrumental Transfer and Alcohol Consumption in Young Male Social Drinkers: Behavioral, Neural and Polygenic Correlates. *Journal of Clinical Medicine*, *8*(8), 1188. https://doi.org/10.3390/jcm8081188

Garbusow, M., Schad, D. J., Sebold, M., Friedel, E., Bernhardt, N., Koch, S. P., Steinacher, B., Kathmann, N., Geurts, D. E., Sommer, C., Muller, D. K., Nebe, S., Paul, S., Wittchen, H. U., Zimmermann, U. S., Walter, H., Smolka, M. N., Sterzer, P., Rapp, M. A., Huys, Q. J., Schlagenhauf, F., & Heinz, A. (2016). Pavlovian-to-instrumental transfer effects in the nucleus accumbens relate to relapse in alcohol dependence. *Addiction Biology*, *21*(3), 719-731. https://doi.org/10.1111/adb.12243

Garbusow, M., Schad, D. J., Sommer, C., Junger, E., Sebold, M., Friedel, E., Wendt, J., Kathmann, N., Schlagenhauf, F., Zimmermann, U. S., Heinz, A., Huys, Q. J., & Rapp, M. A. (2014). Pavlovian-to-instrumental transfer in alcohol dependence: a pilot study. *Neuropsychobiology*, *70*(2), 111-121. https://doi.org/10.1159/000363507

Garofalo, S., & Robbins, T. W. (2017). Triggering Avoidance: Dissociable Influences of Aversive Pavlovian Conditioned Stimuli on Human Instrumental Behavior. *Frontiers in Behavioral Neuroscience*, *11*, 63. https://doi.org/10.3389/fnbeh.2017.00063

Gershman, S. J., Guitart-Masip, M., & Cavanagh, J. F. (2021). Neural signatures of arbitration between Pavlovian and instrumental action selection. *PLOS Computational Biology*, *17*(2), e1008553. https://doi.org/10.1371/journal.pcbi.1008553

Geurts, D. E., Huys, Q. J., den Ouden, H. E., & Cools, R. (2013). Aversive Pavlovian control of instrumental behavior in humans. *Journal of Cognitive Neuroscience*, *25*(9), 1428-1441. https://doi.org/10.1162/jocn_a_00425

Giedd, J. N., Blumenthal, J., Jeffries, N. O., Castellanos, F. X., Liu, H., Zijdenbos, A., Paus, T., Evans, A. C., & Rapoport, J. L. (1999). Brain development during childhood and adolescence: a longitudinal MRI study. *Nature Neuroscience*, *2*(10), 861-863. https://doi.org/10.1038/13158

Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a psychiatric symptom dimension related to deficits in goal-directed control. *Elife*, *5*, e11305. https://doi.org/10.7554/eLife.11305

Gillan, C. M., Otto, A. R., Phelps, E. A., & Daw, N. D. (2015). Model-based learning protects against forming habits. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(3), 523-536. https://doi.org/10.3758/s13415-015-0347-6

Gillan, C. M., & Rutledge, R. B. (2021). Smartphones and the Neuroscience of Mental Health. *Annual Review of Neuroscience*, *44*, 129-151. https://doi.org/10.1146/annurev-neuro-101220-014053

Glantz, M. D., Bharat, C., Degenhardt, L., Sampson, N. A., Scott, K. M., Lim, C. C., Al-Hamzawi, A., Alonso, J., Andrade, L. H., & Cardoso, G. (2020). The epidemiology of alcohol use disorders cross-nationally: Findings from the World Mental Health Surveys. *Addictive Behaviors*, *102*, 106128. https://doi.org/10.1016/j.addbeh.2019.106128

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*(4), 585-595. https://doi.org/10.1016/j.neuron.2010.04.016

Gmel, G., Kuntsche, E., & Rehm, J. (2011). Risky single-occasion drinking: bingeing is not bingeing. *Addiction*, *106*(6), 1037-1045. https://doi.org/10.1111/j.1360-0443.2010.03167.x

Gogtay, N., Giedd, J. N., Lusk, L., Hayashi, K. M., Greenstein, D., Vaituzis, A. C., Nugent, T. F., Herman, D. H., Clasen, L. S., Toga, A. W., Rapoport, J. L., & Thompson, P. M. (2004). Dynamic mapping of human cortical development during childhood through early adulthood. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(21), 8174-8179. https://doi.org/10.1073/pnas.0402680101

Gordon, E. M., Laumann, T. O., Adeyemo, B., Huckins, J. F., Kelley, W. M., & Petersen, S. E. (2016). Generation and Evaluation of a Cortical Area Parcellation from Resting-State Correlations. *Cerebral Cortex*, *26*(1), 288-303. https://doi.org/10.1093/cercor/bhu239

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics*, *5*, 13. https://doi.org/10.3389/fninf.2011.00013

Grüsser, S. M., Wrase, J., Klein, S., Hermann, D., Smolka, M. N., Ruf, M., Weber-Fahr, W., Flor, H., Mann, K., & Braus, D. F. (2004). Cue-induced activation of the striatum and medial prefrontal

cortex is associated with subsequent relapse in abstinent alcoholics. *Psychopharmacology*, *175*(3), 296-302. https://doi.org/10.1007/s00213-004-1828-4

Guitart-Masip, M., Duzel, E., Dolan, R., & Dayan, P. (2014). Action versus valence in decision making. *Trends in Cognitive Sciences*, *18*(4), 194-202. https://doi.org/10.1016/j.tics.2014.01.003

Guitart-Masip, M., Nuys, Q. J. M., Fuentemilla, L., Dayan, P., Duzel, E., & Dolan, R. J. (2012). Go and no-go learning in reward and punishment: Interactions between affect and effect. *Neuroimage*, *62*(1), 154-166. https://doi.org/10.1016/j.neuroimage.2012.04.024

Haber, S. N. (2016). Corticostriatal circuitry. *Dialogues in Clinical Neuroscience*, *18*(1), 7-21. https://www.ncbi.nlm.nih.gov/pubmed/27069376

Hasin, D. S., O'Brien, C. P., Auriacombe, M., Borges, G., Bucholz, K., Budney, A., Compton, W. M., Crowley, T., Ling, W., & Petry, N. M. (2013). DSM-5 Criteria for Substance Use Disorders: Recommendations and Rationale. *American Journal of Psychiatry*, *170*(8), 834-851. https://doi.org/10.1176/appi.ajp.2013.12060782

Heatherton, T. F., Kozlowski, L. T., Frecker, R. C., & Fagerstrom, K. O. (1991). The Fagerstrom Test for Nicotine Dependence - a Revision of the Fagerstrom Tolerance Questionnaire. *British Journal of Addiction*, *86*(9), 1119-1127. https://doi.org/10.1111/j.1360-0443.1991.tb01879.x

Hedge, C., Bompas, A., & Sumner, P. (2020). Task Reliability Considerations in Computational Psychiatry. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *5*(9), 837-839. https://doi.org/10.1016/j.bpsc.2020.05.004

Heilig, M., Epstein, D. H., Nader, M. A., & Shaham, Y. (2016). Time to connect: bringing social context into addiction neuroscience. *Nature Reviews Neuroscience*, *17*(9), 592-599. https://doi.org/10.1038/nrn.2016.67

Heinz, A. (2002). Dopaminergic dysfunction in alcoholism and schizophrenia–psychopathological and behavioral correlates. *European Psychiatry*, *17*(1), 9-16. https://doi.org/10.1016/s0924-9338(02)00628-4

Heinz, A., Kiefer, F., Smolka, M. N., Endrass, T., Beste, C., Beck, A., Liu, S., Genauck, A., Romund, L., & Banaschewski, T. (2020). Addiction Research Consortium: Losing and regaining control over drug intake (ReCoDe)—From trajectories to mechanisms and interventions. *Addiction Biology*, *25*(2), e12866. https://doi.org/10.1111/adb.12866

Heinz, A. J., Beck, A., Meyer-Lindenberg, A., Sterzer, P., & Heinz, A. (2011). Cognitive and neurobiological mechanisms of alcohol-related aggression. *Nature Reviews Neuroscience*, *12*(7), 400. https://doi.org/10.1038/nrn3042

Heng, L.-J., Markham, J. A., Hu, X.-T., & Tseng, K. Y. (2011). Concurrent upregulation of postsynaptic L-type Ca2+ channel function and protein kinase A signaling is required for the periadolescent facilitation of Ca2+ plateau potentials and dopamine D1 receptor modulation in the prefrontal cortex. *Neuropharmacology*, *60*(6), 953-962. https://doi.org/10.1016/j.neuropharm.2011.01.041

Hitchcock, P., Niv, Y., Radulescu, A., & Sims, C. R. (2017). Translating a Reinforcement Learning Task into a Computational Psychiatry Assay: Challenges and Strategies. CogSci,

Hitchcock, P. F., Fried, E. I., & Frank, M. J. (2022). Computational Psychiatry Needs Time and Context. *Annual Review of Psychology*, *73*, 243-270. https://doi.org/10.1146/annurev-psych-021621-124910

Hogarth, L. (2012). Goal-directed and transfer-cue-elicited drug-seeking are dissociated by pharmacotherapy: evidence for independent additive controllers. *Journal of Experimental Psychology: Animal Behavior Processes*, *38*(3), 266-278. https://doi.org/10.1037/a0028914

Hogarth, L. (2022). The Persistence of Addiction is better Explained by Socioeconomic Deprivation-Related Factors Powerfully Motivating Goal-Directed Drug Choice than by Automaticity, Habit or Compulsion Theories Favored by the Brain Disease Model. In *Evaluating the Brain Disease Model of Addiction* (pp. 216-236). Routledge.

Hogarth, L., Balleine, B. W., Corbit, L. H., & Killcross, S. (2012). Associative learning mechanisms underpinning the transition from recreational drug use to addiction. *Annals of the New York Academy of Sciences*, *1282*, 12–24. https://doi.org/10.1111/j.1749-6632.2012.06768.x

Hogarth, L., & Chase, H. W. (2011). Parallel goal-directed and habitual control of human drug-seeking: implications for dependence vulnerability. *Journal of Experimental Psychology: Animal Behavior Processes*, *37*(3), 261-276. https://doi.org/10.1037/a0022913

Hogarth, L., Dickinson, A., Wright, A., Kouvaraki, M., & Duka, T. (2007). The role of drug expectancy in the control of human drug seeking. *Journal of Experimental Psychology: Animal Behavior Processes*, *33*(4), 484-496. https://doi.org/10.1037/0097-7403.33.4.484

Holland, P. C. (2004). Relations between Pavlovian-instrumental transfer and reinforcer devaluation. *Journal of Experimental Psychology: Animal Behavior Processes*, *30*(2), 104. https://doi.org/10.1037/0097-7403.30.2.104

Holmes, N. M., Marchand, A. R., & Coutureau, E. (2010). Pavlovian to instrumental transfer: a neurobehavioural perspective. *Neuroscience & Biobehavioral Reviews*, *34*(8), 1277-1295. https://doi.org/10.1016/j.neubiorev.2010.03.007

Holmes, T. H., & Rahe, R. H. (1967). The social readjustment rating scale. *Journal of Psychosomatic Research*. https://doi.org/10.1016/0022-3999(67)90010-4

Hommel, B. (2011). The Simon effect as tool and heuristic. *Acta Psychologica*, *136*(2), 189-202. https://doi.org/10.1016/j.actpsy.2010.04.011

Hung, Y., Gaillard, S. L., Yarmak, P., & Arsalidou, M. (2018). Dissociations of cognitive inhibition, response inhibition, and emotional interference: Voxelwise ALE meta-analyses of fMRI studies. *Human Brain Mapping*, *39*(10), 4065-4082. https://doi.org/10.1002/hbm.24232

Huppé‐Gourgues, F., & O'donnell, P. (2012). D1–NMDA receptor interactions in the rat nucleus accumbens change during adolescence. *Synapse*, *66*(7), 584-591. https://doi.org/10.1002/syn.21544

Huys, Q. J., Cools, R., Golzer, M., Friedel, E., Heinz, A., Dolan, R. J., & Dayan, P. (2011). Disentangling the roles of approach, activation and valence in instrumental and pavlovian responding. *PLOS Computational Biology*, *7*(4), e1002028. https://doi.org/10.1371/journal.pcbi.1002028

Huys, Q. J. M., Golzer, M., Friedel, E., Heinz, A., Cools, R., Dayan, P., & Dolan, R. J. (2016). The specificity of Pavlovian regulation is associated with recovery from depression. *Psychological Medicine*, *46*(5), 1027-1035. https://doi.org/10.1017/S0033291715002597

Jackson, K. M., Sher, K. J., & Schulenberg, J. E. (2008). Conjoint developmental trajectories of young adult substance use. *Alcoholism: Clinical and Experimental Research*, *32*(5), 723-737. https://doi.org/10.1111/j.1530-0277.2008.00643.x

Jacobi, F., Mack, S., Gerschler, A., Scholl, L., Hofler, M., Siegert, J., Burkner, A., Preiss, S., Spitzer, K., Busch, M., Hapke, U., Gaebel, W., Maier, W., Wagner, M., Zielasek, J., & Wittchen, H. U. (2013). The design and methods of the mental health module in the German Health Interview and Examination Survey for Adults (DEGS1-MH). *International Journal of Methods in Psychiatric Research*, *22*(2), 83-99. https://doi.org/10.1002/mpr.1387

Janssen, F., El Gewily, S., Bardoutsos, A., & Trias-Llimos, S. (2020). Past and Future Alcohol-Attributable Mortality in Europe. *International Journal of Environmental Research and Public Health*, *17*(23), 9024. https://doi.org/10.3390/ijerph17239024

JASP Team. (2021). *JASP (Version 0.16) [Computer software]*. In https://jasp-stats.org/

Jennison, K. M. (2004). The short-term effects and unintended long-term consequences of binge drinking in college: a 10-year follow-up study. *The American Journal of Drug and Alcohol Abuse*, *30*(3), 659-684. https://doi.org/10.1081/ada-200032331

Jones, S. A., Lueras, J. M., & Nagel, B. J. (2018). Effects of Binge Drinking on the Developing Brain Studies in Humans. *Alcohol Research: Current Reviews*, *39*(1), 87-96.

Jupp, B., & Dalley, J. W. (2014). Behavioral endophenotypes of drug addiction: Etiological insights from neuroimaging studies. *Neuropharmacology*, *76 Pt B*, 487-497. https://doi.org/10.1016/j.neuropharm.2013.05.041

Kendler, K. S., Ohlsson, H., Karriker-Jaffe, K. J., Sundquist, J., & Sundquist, K. (2017). Social and economic consequences of alcohol use disorder: a longitudinal cohort and co-relative analysis. *Psychological Medicine*, *47*(5), 925-935. https://doi.org/10.1017/S0033291716003032

Konova, A. B., Lopez-Guzman, S., Urmanche, A., Ross, S., Louie, K., Rotrosen, J., & Glimcher, P. W. (2020). Computational Markers of Risky Decision-making for Identification of Temporal Windows of Vulnerability to Opioid Use in a Real-world Clinical Setting. *JAMA Psychiatry*, *77*(4), 368-377. https://doi.org/10.1001/jamapsychiatry.2019.4013

Koob, G. F., & Le Moal, M. (2005). Plasticity of reward neurocircuitry and the 'dark side' of drug addiction. *Nature Neuroscience*, *8*(11), 1442-1444. https://doi.org/10.1038/nn1105-1442

Koob, G. F., & Volkow, N. D. (2010). Neurocircuitry of addiction. *Neuropsychopharmacology*, *35*(1), 217-238. https://doi.org/10.1038/npp.2009.110

Koob, G. F., & Volkow, N. D. (2016). Neurobiology of addiction: a neurocircuitry analysis. *Lancet Psychiatry*, *3*(8), 760-773. https://doi.org/10.1016/S2215-0366(16)00104-8

Kool, W., Cushman, F. A., & Gershman, S. J. (2016). When does model-based control pay off? *PLOS Computational Biology*, *12*(8). https://doi.org/10.1371/journal.pcbi.1005090

Kool, W., Gershman, S. J., & Cushman, F. A. (2017). Cost-Benefit Arbitration Between Multiple Reinforcement-Learning Systems. *Psychological Science*, *28*(9), 1321-1333. https://doi.org/10.1177/0956797617708288

Kool, W., Gershman, S. J., & Cushman, F. A. (2018). Planning Complexity Registers as a Cost in Metacontrol. *Journal of Cognitive Neuroscience*, *30*(10), 1391-1404. https://doi.org/10.1162/jocn_a_01263

Kouneiher, F., Charron, S., & Koechlin, E. (2009). Motivation and cognitive control in the human prefrontal cortex. *Nature Neuroscience*, *12*(7), 939-945. https://doi.org/10.1038/nn.2321

Kraus, L., Seitz, N. N., Shield, K. D., Gmel, G., & Rehm, J. (2019). Quantifying harms to others due to alcohol consumption in Germany: a register-based study. *BMC Medicine*, *17*(1), 1-9. https://doi.org/10.1186/s12916-019-1290-0

Kühn, S., Witt, C., Banaschewski, T., Barbot, A., Barker, G. J., Büchel, C., Conrod, P. J., Flor, H., Garavan, H., & Ittermann, B. (2016). From mother to child: orbitofrontal cortex gyrification and changes of drinking behaviour during adolescence. *Addiction Biology*, *21*(3), 700–708. https://doi.org/10.1111/adb.12240

Kuitunen-Paul, S., Pfab, S., Garbusow, M., Heinz, A., Kuitunen, P. T., Manthey, J., Nebe, S., Smolka, M. N., & Wittchen, H. U. (2018). Identification of heavy drinking in the 10-item AUDIT: Results from a prospective study among 18-21 years old non-dependent German males. *Journal of Substance Abuse Treatment*, *86*, 94-101. https://doi.org/10.1016/j.jsat.2017.12.011

Kvamme, T. L., Pedersen, M. U., Overgaard, M., Thomsen, K. R., & Voon, V. (2019). Pupillary reactivity to alcohol cues as a predictive biomarker of alcohol relapse following treatment in a pilot study. *Psychopharmacology*, *236*(4), 1233-1243. https://doi.org/10.1007/s00213-018-5131-1

Lannoy, S., Billieux, J., Dormal, V., & Maurage, P. (2019). Behavioral and Cerebral Impairments Associated with Binge Drinking in Youth: A Critical Review. *Psychologica Belgica*, *59*(1), 116-155. https://doi.org/10.5334/pb.476

Lannoy, S., D'Hondt, F., Dormal, V., Billieux, J., & Maurage, P. (2017). Electrophysiological correlates of performance monitoring in binge drinking: Impaired error-related but preserved feedback

processing. *Clinical Neurophysiology*, *128*(11), 2110-2121. https://doi.org/10.1016/j.clinph.2017.08.005

Lees, B., Mewton, L., Stapinski, L. A., Squeglia, L. M., Rae, C. D., & Teesson, M. (2019). Neurobiological and Cognitive Profile of Young Binge Drinkers: a Systematic Review and Meta-Analysis. *Neuropsychology Review*, *29*(3), 357-385. https://doi.org/10.1007/s11065-019-09411-w

Lehrl, S. (2005). Mehrfachwahl-Wortschatz-Intelligenztest MWT-B 5th Edn. *Balingen: Spitta*.

Lewis, A. H., Niznikiewicz, M. A., Delamater, A. R., & Delgado, M. R. (2013). Avoidance-based human Pavlovian-to-instrumental transfer. *European Journal of Neuroscience*, *38*(12), 3740-3748. https://doi.org/10.1111/ejn.12377

Linden-Carmichael, A. N., Vasilenko, S. A., Lanza, S. T., & Maggs, J. L. (2017). High-intensity drinking versus heavy episodic drinking: Prevalence rates and relative odds of alcohol use disorder across adulthood. *Alcoholism: Clinical and Experimental Research*, *41*(10), 1754-1759. https://doi.org/10.1111/acer.13475

Liu, X., Hairston, J., Schrier, M., & Fan, J. (2011). Common and distinct networks underlying reward valence and processing stages: a meta-analysis of functional neuroimaging studies. *Neuroscience & Biobehavioral Reviews*, *35*(5), 1219-1236. https://doi.org/10.1016/j.neubiorev.2010.12.012

Macleod, C. M. (1992). The Stroop Task - the Gold Standard of Attentional Measures. *Journal of Experimental Psychology: General*, *121*(1), 12-14. https://doi.org/10.1037/0096-3445.121.1.12

Madhyastha, T., Peverill, M., Koh, N., McCabe, C., Flournoy, J., Mills, K., King, K., Pfeifer, J., & McLaughlin, K. A. (2018). Current methods and limitations for longitudinal fMRI analysis across development. *Developmental Cognitive Neuroscience*, *33*, 118-128. https://doi.org/10.1016/j.dcn.2017.11.006

Mahlberg, J., Seabrooke, T., Weidemann, G., Hogarth, L., Mitchell, C. J., & Moustafa, A. A. (2021). Human appetitive Pavlovian-to-instrumental transfer: a goal-directed account. *Psychological Research*, *85*(2), 449-463. https://doi.org/10.1007/s00426-019-01266-3

Mahler, S. V., & Berridge, K. C. (2012). What and when to "want"? Amygdala-based focusing of incentive salience upon sugar and sex. *Psychopharmacology*, *221*(3), 407-426. https://doi.org/10.1007/s00213-011-2588-6

Mann, K., & Ackermann, K. (2000). Die OCDS-G: Psychometrische Kennwerte der deutschen Version der obsessive compulsive drinking scale. *Sucht*, *46*(2), 90-100.

Mann, R. E., Sobell, L. C., Sobell, M. B., & Pavan, D. (1985). Reliability of a family tree questionnaire for assessing family history of alcohol problems. *Drug and Alcohol Dependence*, *15*(1-2), 61-67. https://doi.org/10.1016/0376-8716(85)90030-4

Melnikoff, D. E., & Bargh, J. A. (2018). The Mythical Number Two. *Trends in Cognitive Sciences*, *22*(4), 280-293. https://doi.org/10.1016/j.tics.2018.02.001

Mendelsohn, A., Pine, A., & Schiller, D. (2014). Between thoughts and actions: motivationally salient cues invigorate mental action in the human brain. *Neuron*, *81*(1), 207-217. https://doi.org/10.1016/j.neuron.2013.10.019

Meule, A., Vögele, C., & Kübler, A. (2011). Psychometrische evaluation der deutschen Barratt impulsiveness scale–Kurzversion (BIS-15). *Diagnostica*.

Miller, K. J., Shenhav, A., & Ludvig, E. A. (2019). Habits Without Values. *Psychological Review*, *126*(2), 292-311. https://doi.org/10.1037/rev0000120

Miller, S. (2018). *The ASAM principles of addiction medicine*. Lippincott Williams & Wilkins.

Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N. D., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, *1*(9), 680-692. https://doi.org/10.1038/s41562-017-0180-8

Moran, P., Leese, M., Lee, T., Walters, P., Thornicroft, G., & Mann, A. (2003). Standardised Assessment of Personality–Abbreviated Scale (SAPAS): preliminary validation of a brief screen for personality disorder. *The British Journal of Psychiatry*, *183*(3), 228-232. https://doi.org/10.1192/bjp.183.3.228

Murschall, A., & Hauber, W. (2006). Inactivation of the ventral tegmental area abolished the general excitatory influence of Pavlovian cues on instrumental performance. *Learning & Memory*, *13*(2), 123-126. https://doi.org/10.1101/lm.127106

Muthen, B. O., & Muthen, L. K. (2000). The development of heavy drinking and alcohol-related problems from ages 18 to 37 in a US national sample. *Journal of Studies on Alcohol*, *61*(2), 290-300. https://doi.org/10.15288/jsa.2000.61.290

Nadler, N., Delgado, M. R., & Delamater, A. R. (2011). Pavlovian to instrumental transfer of control in a human learning task. *Emotion*, *11*(5), 1112-1123. https://doi.org/10.1037/a0022760

Naimi, T. S., Nelson, D. E., & Brewer, R. D. (2010). The Intensity of Binge Alcohol Consumption Among US Adults. *American Journal of Preventive Medicine*, *38*(2), 201-207. https://doi.org/10.1016/j.amepre.2009.09.039

Nebe, S., Kroemer, N. B., Schad, D. J., Bernhardt, N., Sebold, M., Muller, D. K., Scholl, L., Kuitunen-Paul, S., Heinz, A., Rapp, M. A., Huys, Q. J. M., & Smolka, M. N. (2018). No association of goal-directed and habitual control with alcohol consumption in young adults. *Addiction Biology*, *23*(1), 379-393. https://doi.org/10.1111/adb.12490

Nielsen, F. A., & Hansen, L. K. (2002). Automatic anatomical labeling of Talairach coordinates and generation of volumes of interest via the BrainMap database. *Neuroimage*, *16*(2), 2-6.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139-154. https://doi.org/10.1016/j.jmp.2008.12.005

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424-430. https://doi.org/10.1016/j.tics.2006.07.005

O'Doherty, J. P., Lee, S., Tadayonnejad, R., Cockburn, J., Iigaya, K., & Charpentier, C. J. (2021). Why and how the brain weights contributions from a mixture of experts. *Neuroscience & Biobehavioral Reviews*. https://doi.org/10.1016/j.neubiorev.2020.10.022

O'Doherty, J. P., Lee, S. W., & McNamee, D. (2015). The structure of reinforcement-learning mechanisms in the human brain. *Current Opinion in Behavioral Sciences*, *1*, 94-100. https://doi.org/10.1016/j.cobeha.2014.10.004

O'Rourke, H. P., Fine, K. L., Grimm, K. J., & MacKinnon, D. P. (2021). The Importance of Time Metric Precision When Implementing Bivariate Latent Change Score Models. *Multivariate Behavioral Research*, 1-19. https://doi.org/10.1080/00273171.2021.1874261

Oravecz, Z., & Muth, C. (2018). Fitting growth curve models in the Bayesian framework. *Psychonomic Bulletin & Review*, *25*(1), 235-255. https://doi.org/10.3758/s13423-017-1281-0

Ostlund, S. B., & Balleine, B. W. (2008). On habits and addiction: An associative analysis of compulsive drug seeking. *Drug Discovery Today: Disease Models*, *5*(4), 235-245. https://doi.org/10.1016/j.ddmod.2009.07.004

Pabst, A., & Kraus, L. (2008). Alkoholkonsum, alkoholbezogene Störungen und Trends. Ergebnisse des Epidemiologischen Suchtsurveys 2006. *Sucht*, *54*(7), 36-46. https://doi.org/10.1463/2008.07.05

Paredes-Olay, C., Abad, M. J., Gamez, M., & Rosas, J. M. (2002). Transfer of control between causal predictive judgments and instrumental responding. *Animal Learning & Behavior*, *30*(3), 239-248. https://doi.org/10.3758/bf03192833

Patrick, M. E. (2016). A Call for Research on High-Intensity Alcohol Use. *Alcoholism: Clinical and Experimental Research*, *40*(2), 256-259. https://doi.org/10.1111/acer.12945

Patrick, M. E., Schulenberg, J. E., Martz, M. E., Maggs, J. L., O'Malley, P. M., & Johnston, L. D. (2013). Extreme Binge Drinking Among 12th-Grade Students in the United States Prevalence and Predictors. *JAMA Pediatrics*, *167*(11), 1019-1025. https://doi.org/10.1001/jamapediatrics.2013.2392

Patton, J. H., Stanford, M. S., & Barratt, E. S. (1995). Factor structure of the Barratt impulsiveness scale. *Journal of Clinical Psychology*, *51*(6), 768-774. https://doi.org/10.1002/1097-4679(199511)51:6<768::aid-jclp2270510607>3.0.co;2-1

Pavlov, I. P. (1960). Conditioned reflex: An investigation of the physiological activity of the cerebral cortex.

Peciña, S., & Berridge, K. C. (2013). Dopamine or opioid stimulation of nucleus accumbens similarly amplify cue-triggered 'wanting'for reward: entire core and medial shell mapped as substrates for PIT enhancement. *European Journal of Neuroscience*, *37*(9), 1529-1540. https://doi.org/10.1111/ejn.12174

Penny, W. D., Stephan, K. E., Daunizeau, J., Rosa, M. J., Friston, K. J., Schofield, T. M., & Leff, A. P. (2010). Comparing families of dynamic causal models. *PLOS Computational Biology*, *6*(3), e1000709. https://doi.org/10.1371/journal.pcbi.1000709

Peters, S. K., Dunlop, K., & Downar, J. (2016). Cortico-Striatal-Thalamic Loop Circuits of the Salience Network: A Central Pathway in Psychiatric Disease and Treatment. *Frontiers in Systems Neuroscience*, *10*, 104. https://doi.org/10.3389/fnsys.2016.00104

Peterson, R. A., & Brown, S. P. (2005). On the use of beta coefficients in meta-analysis. *Journal of Applied Psychology*, *90*(1), 175-181. https://doi.org/10.1037/0021-9010.90.1.175

Pettigrew, T. F., & Meertens, R. W. (1995). Subtle and blatant prejudice in Western Europe. *European Journal of Social Psychology*, *25*(1), 57-75. https://doi.org/10.1002/ejsp.2420250106

Prevost, C., Liljeholm, M., Tyszka, J. M., & O'Doherty, J. P. (2012). Neural correlates of specific and general Pavlovian-to-Instrumental Transfer within human amygdalar subregions: a high-resolution fMRI study. *Journal of Neuroscience*, *32*(24), 8383-8390. https://doi.org/10.1523/JNEUROSCI.6237-11.2012

Quail, S. L., Morris, R. W., & Balleine, B. W. (2017). Stress associated changes in Pavlovian-instrumental transfer in humans. *Quarterly Journal of Experimental Psychology*, *70*(4), 675-685. https://doi.org/10.1080/17470218.2016.1149198

Ram, N., & Diehl, M. (2014). Multiple-time-scale design and analysis: Pushing toward real-time modeling of complex developmental processes. In *Handbook of intraindividual variability across the life span* (pp. 328-343). Routledge.

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. (2016). Diffusion Decision Model: Current Issues and History. *Trends in Cognitive Sciences*, *20*(4), 260-281. https://doi.org/10.1016/j.tics.2016.01.007

Redish, A. D., Jensen, S., & Johnson, A. (2008). A unified framework for addiction: vulnerabilities in the decision process. *Behavioral and Brain Sciences*, *31*(4), 415-437; discussion 437-487. https://doi.org/10.1017/S0140525X0800472X

Rehm, J., Gmel Sr, G. E., Gmel, G., Hasan, O. S., Imtiaz, S., Popova, S., Probst, C., Roerecke, M., Room, R., & Samokhvalov, A. V. (2017). The relationship between different dimensions of alcohol use and the burden of disease—an update. *Addiction*, *112*(6), 968-1001. https://doi.org/10.1111/add.13757

Reitan, R. (1979). Trail-Making Test 1979. *Arizona: Reitan Neuropsychology Laboratory*.

Reiter, A. M., Deserno, L., Kallert, T., Heinze, H. J., Heinz, A., & Schlagenhauf, F. (2016a). Behavioral and Neural Signatures of Reduced Updating of Alternative Options in Alcohol-Dependent Patients during Flexible Decision-Making. *Journal of Neuroscience*, *36*(43), 10935-10948. https://doi.org/10.1523/JNEUROSCI.4322-15.2016

Reiter, A. M. F., Deserno, L., Wilbertz, T., Heinze, H. J., & Schlagenhauf, F. (2016b). Risk Factors for Addiction and Their Association with Model-Based Behavioral Control. *Frontiers in Behavioral Neuroscience*, *10*, 26. https://doi.org/10.3389/fnbeh.2016.00026

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., & Nieuwenhuis, S. (2004). The role of the medial frontal cortex in cognitive control. *Science*, *306*(5695), 443-447. https://doi.org/10.1126/science.1100301

Robbins, T. W., & Everitt, B. J. (1999). Drug addiction: bad habits add up. *Nature*, *398*(6728), 567-570. https://doi.org/10.1038/19208

Robinson, M. J. F., & Berridge, K. C. (2013). Instant Transformation of Learned Repulsion into Motivational "Wanting". *Current Biology*, *23*(4), 282-289. https://doi.org/10.1016/j.cub.2013.01.016

Robinson, T. E., & Berridge, K. C. (1993). The neural basis of drug craving: an incentive-sensitization theory of addiction. *Brain Research Reviews*, *18*(3), 247-291. https://doi.org/10.1016/0165-0173(93)90013-p

Robinson, T. E., & Berridge, K. C. (2008). The incentive sensitization theory of addiction: some current issues. *Philosophical Transactions of the Royal Society B-Biological Sciences*, *363*(1507), 3137-3146. https://doi.org/10.1098/rstb.2008.0093

Rosas, J. M., Paredes-Olay, M. C., García-Gutiérrez, A., Espinosa, J. J., & Abad, M. J. (2010). Outcome-specific transfer between predictive and instrumental learning is unaffected by extinction but reversed by counterconditioning in human participants. *Learning and Motivation*, *41*(1), 48-66. https://doi.org/10.1016/j.lmot.2009.09.002

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. Version 0.5–12 (BETA). *Journal of Statistical Software*, *48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Rouder, J. N., & Haaf, J. M. (2019). A psychometrics of individual differences in experimental tasks. *Psychonomic Bulletin & Review*, *26*(2), 452-467. https://doi.org/10.3758/s13423-018-1558-y

Rouder, J. N., & Haaf, J. M. (2021). Are There Reliable Qualitative Individual Difference in Cognition? *Journal of Cognition*, *4*(1). https://doi.org/10.5334/joc.131

Saß, H., Wittchen, H.-U., Zaudig, M., & Houben, I. (2003). DSM-IV-TR–Diagnostisches und Statistisches Manual Psychischer Störungen–Textrevision. *Hogrefe, Göttingen*.

Schad, D. J., Garbusow, M., Friedel, E., Sommer, C., Sebold, M., Hagele, C., Bernhardt, N., Nebe, S., Kuitunen-Paul, S., Liu, S., Eichmann, U., Beck, A., Wittchen, H. U., Walter, H., Sterzer, P., Zimmermann, U. S., Smolka, M. N., Schlagenhauf, F., Huys, Q. J. M., Heinz, A., & Rapp, M. A. (2019). Neural correlates of instrumental responding in the context of alcohol-related cues index disorder severity and relapse risk. *European Archives of Psychiatry and Clinical Neuroscience*, *269*(3), 295-308. https://doi.org/10.1007/s00406-017-0860-4

Schad, D. J., Junger, E., Sebold, M., Garbusow, M., Bernhardt, N., Javadi, A. H., Zimmermann, U. S., Smolka, M. N., Heinz, A., Rapp, M. A., & Huys, Q. J. (2014). Processing speed enhances model-based over model-free reinforcement learning in the presence of high working memory functioning. *Frontiers in Psychology*, *5*, 1450. https://doi.org/10.3389/fpsyg.2014.01450

Schönberg, T., Daw, N. D., Joel, D., & O'Doherty, J. P. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *Journal of Neuroscience*, *27*(47), 12860-12867. https://doi.org/10.1523/JNEUROSCI.2496-07.2007

Schulenberg, J. E., & Maggs, J. L. (2002). A developmental perspective on alcohol use and heavy drinking during adolescence and the transition to young adulthood. *Journal of Studies on Alcohol*, *Supplement*(14), 54-70. https://doi.org/10.15288/jsas.2002.s14.54

Schultz, W. (1986). Responses of midbrain dopamine neurons to behavioral trigger stimuli in the monkey. *Journal of Neurophysiology*, *56*(5), 1439-1461. https://doi.org/10.1152/jn.1986.56.5.1439

Schultz, W. (2007a). Behavioral dopamine signals. *Trends in Neurosciences*, *30*(5), 203-210. https://doi.org/10.1016/j.tins.2007.03.007

Schultz, W. (2007b). Multiple dopamine functions at different time courses. *Annual Review of Neuroscience*, *30*, 259-288. https://doi.org/10.1146/annurev.neuro.28.061604.135722

Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience*, *13*(3), 900-913. https://www.ncbi.nlm.nih.gov/pubmed/8441015

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593-1599. https://doi.org/10.1126/science.275.5306.1593

Schultz, W., Tremblay, L., & Hollerman, J. R. (2000). Reward processing in primate orbitofrontal cortex and basal ganglia. *Cerebral Cortex*, *10*(3), 272-284. https://doi.org/10.1093/cercor/10.3.272

Schwöbel, S., Marković, D., Smolka, M. N., & Kiebel, S. J. (2021). Balancing control: a Bayesian interpretation of habitual and goal-directed behavior. *Journal of Mathematical Psychology*, *100*, 102472. https://doi.org/10.1016/j.bandc.2022.105843

Seabrooke, T., Hogarth, L., Edmunds, C. E. R., & Mitchell, C. J. (2019). Goal-directed control in Pavlovian-instrumental transfer. *Journal of Experimental Psychology: Animal Learning and Cognition*, *45*(1), 95-101. https://doi.org/10.1037/xan0000191

Seabrooke, T., Hogarth, L., & Mitchell, C. J. (2016). The propositional basis of cue-controlled reward seeking. *Quarterly Journal of Experimental Psychology*, *69*(12), 2452-2470. https://doi.org/10.1080/17470218.2015.1115885

Seabrooke, T., Le Pelley, M. E., Hogarth, L., & Mitchell, C. J. (2017). Evidence of a goal-directed process in human Pavlovian-instrumental transfer. *Journal of Experimental Psychology: Animal Learning and Cognition*, *43*(4), 377-387. https://doi.org/10.1037/xan0000147

Seabrooke, T., Le Pelley, M. E., Porter, A., & Mitchell, C. J. (2018). Extinguishing cue-controlled reward choice: Effects of Pavlovian extinction on outcome-selective Pavlovian-instrumental transfer. *Journal of Experimental Psychology: Animal Learning and Cognition*, *44*(3), 280-292. https://doi.org/10.1037/xan0000176

Sebold, M., Deserno, L., Nebe, S., Schad, D. J., Garbusow, M., Hagele, C., Keller, J., Junger, E., Kathmann, N., Smolka, M. N., Rapp, M. A., Schlagenhauf, F., Heinz, A., & Huys, Q. J. (2014). Model-based and model-free decisions in alcohol dependence. *Neuropsychobiology*, *70*(2), 122-131. https://doi.org/10.1159/000362840

Sebold, M., Nebe, S., Garbusow, M., Guggenmos, M., Schad, D. J., Beck, A., Kuitunen-Paul, S., Sommer, C., Frank, R., Neu, P., Zimmermann, U. S., Rapp, M. A., Smolka, M. N., Huys, Q. J. M., Schlagenhauf, F., & Heinz, A. (2017). When Habits Are Dangerous: Alcohol Expectancies and Habitual Decision Making Predict Relapse in Alcohol Dependence. *Biological Psychiatry*, *82*(11), 847-856. https://doi.org/10.1016/j.biopsych.2017.04.019

Sebold, M., Schad, D. J., Nebe, S., Garbusow, M., Junger, E., Kroemer, N. B., Kathmann, N., Zimmermann, U. S., Smolka, M. N., Rapp, M. A., Heinz, A., & Huys, Q. J. (2016). Don't Think, Just Feel the Music: Individuals with Strong Pavlovian-to-Instrumental Transfer Effects Rely Less on Model-based Reinforcement Learning. *Journal of Cognitive Neuroscience*, *28*(7), 985-995. https://doi.org/10.1162/jocn_a_00945

Shahar, N., Hauser, T. U., Moutoussis, M., Moran, R., Keramati, M., consortium, N., & Dolan, R. J. (2019a). Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. *PLOS Computational Biology*, *15*(2), e1006803. https://doi.org/10.1371/journal.pcbi.1006803

Shahar, N., Moran, R., Hauser, T. U., Kievit, R. A., McNamee, D., Moutoussis, M., Consortium, N., & Dolan, R. J. (2019b). Credit assignment to state-independent task representations and its relationship with model-based decision making. *Proceedings of the National Academy of Sciences*, *116*(32), 15871-15876. https://doi.org/10.1073/pnas.1821647116

Shield, K., Manthey, J., Rylett, M., Probst, C., Wettlaufer, A., Parry, C. D. H., & Rehm, J. (2020). National, regional, and global burdens of disease from 2000 to 2016 attributable to alcohol use: a comparative risk assessment study. *Lancet Public Health*, *5*(1), E51-E61. https://doi.org/10.1016/S2468-2667(19)30231-2

Shield, K. D., Parry, C., & Rehm, J. (2014). Chronic diseases and conditions related to alcohol use. *Alcohol Research: Current Reviews*, *35*(2), 155.

Shulman, E. P., Smith, A. R., Silva, K., Icenogle, G., Duell, N., Chein, J., & Steinberg, L. (2016). The dual systems model: Review, reappraisal, and reaffirmation. *Developmental Cognitive Neuroscience*, *17*, 103-117. https://doi.org/10.1016/j.dcn.2015.12.010

Simon, J. R., & Rudell, A. P. (1967). Auditory S-R compatibility: the effect of an irrelevant cue on information processing. *Journal of Applied Psychology*, *51*(3), 300. https://doi.org/10.1037/h0020586

Skinner, H. A., & Allen, B. A. (1982). Alcohol dependence syndrome: measurement and validation. *Journal of Abnormal Psychology*, *91*(3), 199-209. https://doi.org/10.1037/0021-843x.91.3.199

Soder, H. E., Webber, T. A., Bornovalova, M. A., Park, J. Y., & Potts, G. F. (2019). A test of dopamine hyper- and hyposensitivity in alcohol use. *Addictive Behaviors*, *90*, 395-401. https://doi.org/10.1016/j.addbeh.2018.12.002

Sommer, C., Birkenstock, J., Garbusow, M., Obst, E., Schad, D. J., Bernhardt, N., Huys, Q. M., Wurst, F. M., Weinmann, W., Heinz, A., Smolka, M. N., & Zimmermann, U. S. (2020). Dysfunctional approach behavior triggered by alcohol-unrelated Pavlovian cues predicts long-term relapse in alcohol dependence. *Addiction Biology*, *25*(1), e12703. https://doi.org/10.1111/adb.12703

Sommer, C., Garbusow, M., Junger, E., Pooseh, S., Bernhardt, N., Birkenstock, J., Schad, D. J., Jabs, B., Glockler, T., Huys, Q. M., Heinz, A., Smolka, M. N., & Zimmermann, U. S. (2017). Strong seduction: impulsivity and the impact of contextual cues on instrumental behavior in alcohol dependence. *Translational Psychiatry*, *7*(8), e1183. https://doi.org/10.1038/tp.2017.158

Squeglia, L. M., Pulido, C., Wetherill, R. R., Jacobus, J., Brown, G. G., & Tapert, S. F. (2012). Brain response to working memory over three years of adolescence: influence of initiating heavy drinking. *Journal of Studies on Alcohol and Drugs*, *73*(5), 749-760. https://doi.org/10.15288/jsad.2012.73.749

Stephan, K. E., Penny, W. D., Moran, R. J., den Ouden, H. E., Daunizeau, J., & Friston, K. J. (2010). Ten simple rules for dynamic causal modeling. *Neuroimage*, *49*(4), 3099-3109. https://doi.org/10.1016/j.neuroimage.2009.11.015

Stevens, A. K., Blanchard, B. E., Shi, M., & Littlefield, A. K. (2020). Alcohol Use Disorder: Long‐Term Consequences. *The Wiley Encyclopedia of Health Psychology*, 437-444. https://doi.org/10.1002/9781119057840.ch178

Stockwell, T., Andreasson, S., Cherpitel, C., Chikritzhs, T., Dangardt, F., Holder, H., Naimi, T., & Sherk, A. (2021). The burden of alcohol on health care during COVID-19. *Drug and Alcohol Review*, *40*(1), 3-7. https://doi.org/10.1111/dar.13143

Stockwell, T., Chikritzhs, T., & Dawson, D. (2000). International guide for monitoring alcohol consumption and related harm. *Geneva, Switzerland: World Health Organization*. https://doi.org/10.13140/RG.2.2.29007.48808

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643-662. https://doi.org/10.1037/h0054651

Stuke, H., Gutwinski, S., Wiers, C. E., Schmidt, T. T., Gröpper, S., Parnack, J., Gawron, C., Attar, C. H., Spengler, S., & Walter, H. (2016). To drink or not to drink: Harmful drinking is associated with hyperactivation of reward areas rather than hypoactivation of control areas in men. *Journal of Psychiatry and Neuroscience*, *41*(3), E24-E36. https://doi.org/10.1503/jpn.150203

Sussman, S., & Arnett, J. J. (2014). Emerging Adulthood: Developmental Period Facilitative of the Addictions. *Evaluation & the Health Professions*, *37*(2), 147-155. https://doi.org/10.1177/0163278714521812

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, *3*(1), 9-44. https://doi.org/10.1007/BF00115009

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: an introduction* (Vol. 22447). MIT Press.

Swart, J. C., Frank, M. J., Maatta, J. I., Jensen, O., Cools, R., & den Ouden, H. E. M. (2018). Frontal network dynamics reflect neurocomputational mechanisms for reducing maladaptive biases in motivated action. *PLOS Biology*, *16*(10), e2005979. https://doi.org/10.1371/journal.pbio.2005979

Swart, J. C., Frobose, M. I., Cook, J. L., Geurts, D. E., Frank, M. J., Cools, R., & den Ouden, H. E. (2017). Catecholaminergic challenge uncovers distinct Pavlovian and instrumental mechanisms of motivated (in)action. *Elife*, *6*, e22169. https://doi.org/10.7554/eLife.22169

Talmi, D., Seymour, B., Dayan, P., & Dolan, R. J. (2008). Human pavlovian-instrumental transfer. *Journal of Neuroscience*, *28*(2), 360-368. https://doi.org/10.1523/JNEUROSCI.4028-07.2008

Taylor, R. (1990). Interpretation of the correlation coefficient: a basic review. *Journal of Diagnostic Medical Sonography*, *6*(1), 35-39. https://doi.org/10.1177/875647939000600106

Toyama, A., Katahira, K., & Ohira, H. (2017). A simple computational algorithm of model-based choice preference. *Cognitive, Affective, & Behavioral Neuroscience*, *17*(4), 764-783. https://doi.org/10.3758/s13415-017-0511-2

Toyama, A., Katahira, K., & Ohira, H. (2019). Biases in estimating the balance between model-free and model-based learning systems due to model misspecification. *Journal of Mathematical Psychology*, *91*, 88-102. https://doi.org/10.1016/j.jmp.2019.03.007

Tucker, J. S., Orlando, M., & Ellickson, P. L. (2003). Patterns and correlates of binge drinking trajectories from early adolescence to young adulthood. *Health Psychology*, *22*(1), 79-87. https://doi.org/10.1037//0278-6133.22.1.79

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, *15*(1), 273-289. https://doi.org/10.1006/nimg.2001.0978

Urban, N. B., Kegeles, L. S., Slifstein, M., Xu, X., Martinez, D., Sakr, E., Castillo, F., Moadel, T., O'Malley, S. S., & Krystal, J. H. (2010). Sex differences in striatal dopamine release in young adults after oral alcohol challenge: a positron emission tomography imaging study with [11C] raclopride. *Biological Psychiatry*, *68*(8), 689-696. https://doi.org/10.1016/j.biopsych.2010.06.005

van Steenbergen, H., Watson, P., Wiers, R. W., Hommel, B., & de Wit, S. (2017). Dissociable corticostriatal circuits underlie goal-directed vs. cue-elicited habitual food seeking after satiation: evidence from a multimodal MRI study. *European Journal of Neuroscience*, *46*(2), 1815-1827. https://doi.org/10.1111/ejn.13586

van Timmeren, T., Quail, S. L., Balleine, B. W., Geurts, D. E., Goudriaan, A. E., & van Holst, R. J. (2020). Intact corticostriatal control of goal-directed action in Alcohol Use Disorder: a Pavlovian-to-instrumental transfer and outcome-devaluation study. *Scientific Reports*, *10*(1), 1-12. https://doi.org/10.1038/s41598-020-61892-5

Veer, I. M., Jetzschmann, P., Garbusow, M., Nebe, S., Frank, R., Kuitunen-Paul, S., Sebold, M., Ripke, S., Heinz, A., & Friedel, E. (2019). Nucleus accumbens connectivity at rest is associated with

alcohol consumption in young male adults. *European Neuropsychopharmacology*, *29*(12), 1476-1485. https://doi.org/10.1016/j.euroneuro.2019.10.008

Verhoog, S., Dopmeijer, J. M., de Jonge, J. M., van der Heijde, C. M., Vonk, P., Bovens, R. H. L. M., de Boer, M. R., Hoekstra, T., Kunst, A. E., Wiers, R. W., & Kuipers, M. A. G. (2020). The Use of the Alcohol Use Disorders Identification Test - Consumption as an Indicator of Hazardous Alcohol Use among University Students. *European Addiction Research*, *26*(1), 1-9. https://doi.org/10.1159/000503342

Vollstädt-Klein, S., Loeber, S., Richter, A., Kirsch, M., Bach, P., von der Goltz, C., Hermann, D., Mann, K., & Kiefer, F. (2012). Validating incentive salience with functional magnetic resonance imaging: association between mesolimbic cue reactivity and attentional bias in alcohol-dependent patients. *Addiction Biology*, *17*(4), 807-816. https://doi.org/10.1111/j.1369-1600.2011.00352.x

Voon, V., Derbyshire, K., Ruck, C., Irvine, M. A., Worbe, Y., Enander, J., Schreiber, L. R., Gillan, C., Fineberg, N. A., Sahakian, B. J., Robbins, T. W., Harrison, N. A., Wood, J., Daw, N. D., Dayan, P., Grant, J. E., & Bullmore, E. T. (2015). Disorders of compulsivity: a common bias towards learning habits. *Molecular Psychiatry*, *20*(3), 345-352. https://doi.org/10.1038/mp.2014.44

Waltmann, M., Schlagenhauf, F., & Deserno, L. (2021). Sufficient Reliability of the Behavioral and Computational Read-Outs of a Probabilistic Reversal Learning Task. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01739-7

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., Hassabis, D., & Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, *21*(6), 860-+. https://doi.org/10.1038/s41593-018-0147-8

Watkins, C. J. C. H. (1989). *Learning form delayed rewards*

Watson, P., O'Callaghan, C., Perkes, I., Bradfield, L., & Turner, K. (2022). Making habits measurable beyond what they are not: a focus on associative dual-process models. *PsyArXiv*. https://doi.org/10.31234/osf.io/7m6c8

Watson, P., Wiers, R. W., Hommel, B., & de Wit, S. (2014). Working for food you don't desire. Cues interfere with goal-directed food-seeking. *Appetite*, *79*, 139-148. https://doi.org/10.1016/j.appet.2014.04.005

Wechsler, D. (1997). *WAIS-III, Wechsler Adult Intelligence Scale: Administration and Scoring Manual.* The Psychological Corporation.

Wetherill, R. R., Squeglia, L. M., Yang, T. T., & Tapert, S. F. (2013). A longitudinal examination of adolescent response inhibition: neural differences before and after the initiation of heavy drinking. *Psychopharmacology*, *230*(4), 663-671. https://doi.org/10.1007/s00213-013-3198-2

White, A. M., Kraus, C. L., & Swartzwelder, H. S. (2006). Many college freshmen drink at levels far beyond the binge threshold. *Alcoholism: Clinical and Experimental Research*, *30*(6), 1006-1010. https://doi.org/10.1111/j.1530-0277.2006.00122.x

Wiers, C. E., Stelzel, C., Park, S. Q., Gawron, C. K., Ludwig, V. U., Gutwinski, S., Heinz, A., Lindenmeyer, J., Wiers, R. W., Walter, H., & Bermpohl, F. (2014). Neural Correlates of Alcohol-Approach Bias in Alcohol Addiction: the Spirit is Willing but the Flesh is Weak for Spirits. *Neuropsychopharmacology*, *39*(3), 688-697. https://doi.org/10.1038/npp.2013.252

Windle, M., Mun, E. Y., & Windle, R. C. (2005). Adolescent-to-young adulthood heavy drinking trajectories and their prospective predictors. *Journal of Studies on Alcohol*, *66*(3), 313-322. https://doi.org/10.15288/jsa.2005.66.313

Wise, R. A. (1987). The role of reward pathways in the development of drug dependence. *Pharmacology & Therapeutics*, *35*(1-2), 227-263. https://doi.org/10.1016/0163-7258(87)90108-2

Wise, R. A. (1988). The neurobiology of craving: implications for the understanding and treatment of addiction. *Journal of Abnormal Psychology*, *97*(2), 118-132. https://doi.org/10.1037//0021-843x.97.2.118

Wittchen, H.-U., & Pfister, H. (1997). DIA-X-Interviews: Manual für Screening-Verfahren und Interview; Interviewheft.

Witteman, J., Post, H., Tarvainen, M., de Bruijn, A., Perna Ede, S., Ramaekers, J. G., & Wiers, R. W. (2015). Cue reactivity and its relation to craving and relapse in alcohol dependence: a combined laboratory and field study. *Psychopharmacology*, *232*(20), 3685-3696. https://doi.org/10.1007/s00213-015-4027-6

Woicik, P. A., Stewart, S. H., Pihl, R. O., & Conrod, P. J. (2009). The substance use risk profile scale: A scale measuring traits linked to reinforcement-specific substance use profiles. *Addictive Behaviors*, *34*(12), 1042-1055. https://doi.org/10.1016/j.addbeh.2009.07.001

World Health Organization. (2019). *Global status report on alcohol and health 2018*. World Health Organization.

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100-1122. https://doi.org/10.1177/1745691617693393

Yee, D. M., & Braver, T. S. (2018). Interactions of Motivation and Cognitive Control. *Current Opinion in Behavioral Sciences*, *19*, 83-90. https://doi.org/10.1016/j.cobeha.2017.11.009

Zdankiewicz-Ścigała, E., & Ścigała, D. K. (2018). Trauma, temperament, alexithymia, and dissociation among persons addicted to alcohol: mediation model of dependencies. *Frontiers in Psychology*, *9*, 1570. https://doi.org/10.3389/fpsyg.2018.01570

Zdankiewicz-Ścigała, E., & Ścigała, D. K. (2020). Attachment style, early childhood trauma, alexithymia, and dissociation among persons addicted to alcohol: Structural equation model of dependencies. *Frontiers in Psychology*, *10*, 2957. https://doi.org/10.3389/fpsyg.2019.02957

Zeidman, P., Jafarian, A., Corbin, N., Seghier, M. L., Razi, A., Price, C. J., & Friston, K. J. (2019a). A guide to group effective connectivity analysis, part 1: First level analysis with DCM for fMRI. *Neuroimage*, *200*, 174-190. https://doi.org/10.1016/j.neuroimage.2019.06.031

Zeidman, P., Jafarian, A., Seghier, M. L., Litvak, V., Cagnan, H., Price, C. J., & Friston, K. J. (2019b). A guide to group effective connectivity analysis, part 2: Second level analysis with PEB. *Neuroimage*, *200*, 12-25. https://doi.org/10.1016/j.neuroimage.2019.06.032

## Acknowledgements

The whole PhD has been a fulfilling journey for me, through which I feel I have understood human behaviors a little bit more than before, and my curiosity is satisfied to a great extent. First and foremost, I am deeply grateful to my PhD supervisor Prof. Michael Smolka for his guidance, invaluable feedback, generous support, and patience throughout the whole journey. I would like to express my deep gratitude to all my collaborators—none of the publications would have been possible if it were not for their inspiring and constructive feedback. I'm also extremely grateful to the DFG for funding the FOR1617 and TRR265 B03 projects, making this whole dissertation possible. Moreover, I could not have undertaken this journey without my colleagues at Section of Systems Neuroscience, not only because of the in-depth scientific exchanges, they also made this place feel like home for me during the last years.

This endeavor would not have been possible without the support from my parents—they always encourage me to pursue my own goals and back me up with unconditional trust and love. Words cannot express my gratitude to Wei—one of the best things that have happened during this journey is meeting you and knowing you.

## Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; Hilfe dritter wurde nur in wissenschaftlich vertretbarem und prüfungsrechtlich zulässigem Ausmaß in Anspruch genommen, die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Es sind keine unzulässigen geldwerten Leistungen, weder unmittelbar noch mittelbar, im Zusammenhang mit dem Inhalt der vorliegenden Dissertation an Dritte erfolgt. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Diese Dissertation wurde in der Zeit vom 05.2017 bis 08.2022 unter der Betreuung von Prof. Dr. med. Michael N. Smolka im Forschungsbereich Systemische Neurowissenschaften, Klinik und Poliklinik für Psychiatrie und Psychotherapie, Technische Universität Dresden angefertigt.

Es haben keine früheren erfolglosen Promotionsverfahren stattgefunden.

Die Promotionsordnung des Bereichs Mathematik und Naturwissenschaften der Technischen Universität Dresden, in der Fassung vom 23.02.2011, letzte Änderung 23.05.2018, wird anerkannt.

Dresden, August 2022

Hao Chen