北京大学

本科生毕业论文

题目：情境对于人类在探索利用任务中的次优性行为的影响

The Description and Experience Gap in Human Exploration and Exploitation Tradeoff

姓　　名：李佳霖

学　　号：2000013713

院　　系：心理与认知科学学院

专　　业：心理学

导师姓名：李健 研究员

二〇二四年六月

版权声明

摘要

探索和利用的权衡是研究智能体如何与环境交互，并在环境中学习和产生适应性行为的重要框架。在传统的决策理论中，研究者发现了一系列人类如何偏离最优决策和产生认知偏差的现象。然而，这些现象均建立在描述性情境中，即与选项有关的信息以符号的形式或信息传递给人类。事实上，已有研究证明人类在经历性情境中也会产生认知偏差，且与在描述性情境中的观察有所差异，这种差异被称为描述和经历情境之间的差异。综上，本研究采用了一种简化的探索和利用范式，以剥离环境无关因素对于决策的影响，单纯观察人们在序列规划问题中描述和经历情境中的差异。结果表明，在两种情境下，人们对环境变量，例如最大奖赏，试次长度，当前奖赏和最大奖赏之间的差距等变量的敏感性不同。此外，本研究还发现在试次水平上，人们会动态地利用之前的信息来调整自己当前的行为表现，以适应环境的变化。基于回归和原始数据层面的观察，本研究提出了一种新的学习启发式模型，用以解释人们如何在探索和利用环境下进行决策，并发现相较于原始模型具有了更好的表现。本研究发现了情境对于人类在探索和利用权衡任务中的影响，并从试次内和试次间水平上区分了人类用于适应环境的方法，将其定义为局部和全局的动态调节。后续研究可以结合规划问题相关的强化学习及启发式认知模型对其进行更深入的研究。

关键词：探索和利用权衡，描述和经历差距，全局和局部调节，启发式决策，适应性学习

ABSTRACT

The Exploration and Exploitation tradeoff is a crucial architecture for studying how intelligent agents interact with their environment, learn, and generate adaptive behaviors. In traditional decision-making theories, researchers have identified a range of phenomena illustrating how humans deviate from optimal decisions and exhibit cognitive biases. However, these phenomena are established within descriptive contexts, where information relevant to choices is presented symbolically to humans. Indeed, research has shown that humans also exhibit cognitive biases in the experience contexts, differing from observations in descriptive contexts—a phenomenon termed the description-experience gap. This study adopts a minimalistic exploration-exploitation paradigm to isolate the impact of environment-irrelevant factors on decision-making and solely observe the description-experience gap in such sequential planning problems. Results indicate varying sensitivities to environmental variables such as maximum reward, trial length, and discrepancy between current and maximum reward(*gap*) across both contexts. Additionally, the study finds people will dynamically utilize prior information at the trial level to adjust current behavior and adapt to environmental changes and stochasticity. Based on regression and raw data observations, a novel learning heuristic model is proposed, explaining how individuals make decisions under exploration and exploitation conditions, showing superior performance compared to the original model. This study reveals the influence of context on humans' exploration-exploitation trade-off. We also propose and differentiates two pivotal concept in human adaptive behavior—local and global dynamic adjustments. Future research could integrate reinforcement learning and heuristic cognitive models related to planning problems for further investigation.

KEY WORDS: Exploration-Exploitation tradeoff, Description-Experience Gap, Local and global adjustment, Heuristic decision making, Adaptive learning.

目 录

第一章 引言

探索和利用的权衡(Exploration-exploitation tradeoff)是研究智能体和环境如何交互，以及智能体如何在环境中进行学习的重要范式(Daw et al, 2006)。事实上，在我们的日常生活中经常会存在探索和利用之间权衡的场景。不妨想象以下的场景：你此时此刻要选择去到一家餐馆去吃饭，你既可以选择去一家你所熟悉口味的餐馆吃饭，也可以选择去一家你之前从未去过的餐馆吃饭。在这种情况下，选择去到一家熟悉的餐馆吃饭可以看作是利用对于环境的已知信息来作出决策的表现，通常情况下利用带来的结果是确定的，但决策可能并不一定是最优的；而选择随机去到一家陌生的餐馆则可以看作是探索环境的表现，这可能带来更大的收益，当然也伴随着一定潜在的风险。

根据强化学习理论的观点，人们在探索和利用动作之间的权衡取决于选项本身能够带来的收益的大小。例如，被试在与环境交互的过程中不断对每个选项进行采样，通过学习的方式更新选项本身的价值大小。其中，学习率便一定程度上反映被试受到选项本身带来收益改变从而更新选项价值的快慢速度。在早期的研究中，研究者便发现了纹状体的多巴胺神经元和预期误差之间存在一定关系(Schultz et al., 1997)。当预期误差较大时，多巴胺神经元的发放幅度也相应较大。随着学习的不断进行，多巴胺神经元的发放幅度减小，且逐渐由在奖赏出现的位置发放传播至能够预测奖赏出现的位置。

事实上，探索和利用之间的权衡除受到选项本身价值的调节外，一部分还与人们当前所处的情境(Context)有关。例如，对比在一个陌生的城市和在一个你所熟悉的城市，你寻找一家餐馆吃饭。在陌生的城市中，由于对于环境的不确定性，你可能需要大致形成对于每一家餐馆大致的印象。而在熟悉的城市中，你对于大部分餐馆的口味如何已经大致有所了解。因此，在两种不同的情境中，人们对于探索和利用的权衡可能出现差异。

在前人的研究中，这种差异被称为描述和经历情境之间的差距(Description-Experience Gap) (Hertwig & Erev, 2009)。在上述选择餐馆的场景中，陌生的城市便可以当作经历情境，即人们需要自己去感受选项可能带来的潜在收益。熟悉的情境便可以当作描述情境，即人们清楚地知道选项背后的结果。

值得注意的是，在早期有关决策的理论中，其基本都是建立在描述情境下的讨论，即人们清楚地知晓并能够计算每一个选择背后带来的期望收益的大小，决策变量（如收益和概率）通过符号明确传达。期望效用理论(Expected utility theory)在公理化假设的基础上，建立了理性人在风险决策下的选择和分析的框架(Von Neumann & Morgenstern, 1947)。前景理论(Prospect theory)的提出解释人们在风险和不确定性决策下的一系列偏差(Kahneman & Tversky, 2013)，包括损失厌恶(Loss aversion)：人们对损失的反应强烈于同等大小的收益；确定性效应(Certain effect)：确定获得某样东西比可能获得更大收益的不确定性事件更有吸引力；反射性效应(Reflect effect)：在面对收益时，人们通常表现出风险规避的倾向，而在面对损失时，则表现出风险寻求的倾向；参照依赖(Reference dependent)：决策是基于对结

果相对于参照点的增益或损失的评估，而不是基于最终状态。

然而，当选择背后的概率收益变得更加不确定时，即人们通过经历(Experience)的方式，与环境交互并得到环境反馈来探索选择本身的价值时，人们的决策行为会受到哪些因素的调控，以及上述在描述情境中人们所表现出的各种认知偏差是否仍然存在目前尚不明确。

在前人探究人们如何在经历情境中进行决策的研究中，研究者往往采用如下**图 1** 所示的实验范式(Wulff et al., 2018)，被试需要通过选择左边或右边的选项来使自己的累计收益最大化。值得注意的是，被试对两个选项背后的期望价值并不知情。因此，被试只能通过点按的方式来对选项的价值进行估计。在具体呈现方式上分为三种，被试可以自主决定采样次数**(图 1A)**，在被试决策后给予选择选项的部分反馈**(图 1B)**或是在决策后对两个选项当前价值都进行反馈**(图 1C)**。此外，由于选项本身每回合的收益都是具有概率性的，相较于描述性情境直接呈现选项的收益和概率而言，往往被试还需要估计选项背后收益的概率。

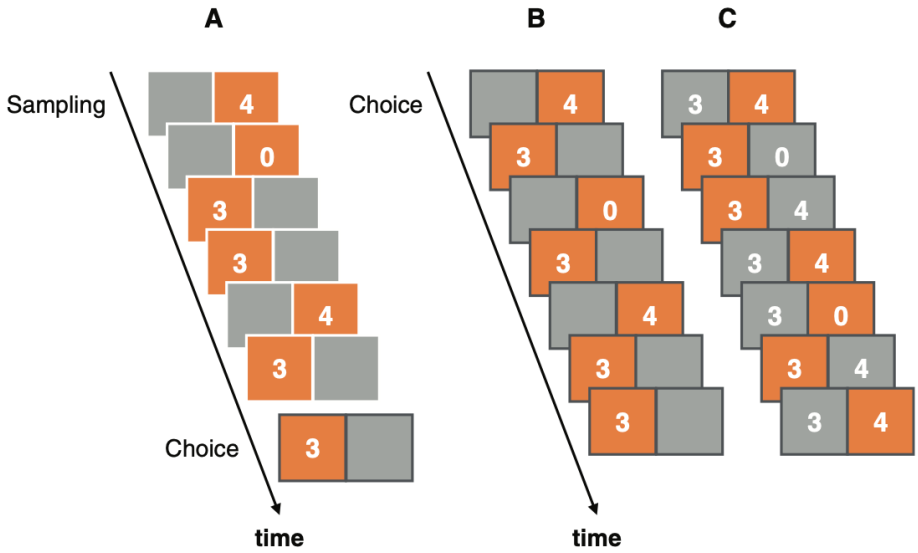

图 1 收录自(Wulff et al., 2018)中探究经历情境决策的采样范式。

A：自主采样：被试在决策前对两个选项进行采样，了解对应的收益和概率后进行决策；

B：部分反馈：每次决策后给被试呈现选择的选项的收益；

C：全反馈：每次决策后给被试同时呈现选择和未选择选项的收益。

Erev 等人(2009)发现，当处于描述性情境下时，人们整体表现为对罕见事件(Rare event)的高估和对频发事件(Frequent event)的低估，即主观概率曲线的反 S 型扭曲。相反，当处于经历性情境时，人们整体表现为对罕见事件的低估和对频发事件的高估，即主观概率的 S 型扭曲。进一步，在收益和损失两种环境下，人们在描述和经验情境间的差距体现的主观概率的扭曲将会展现出不同，如下**图 2** 所示。当人们处于收益环境时，描述情境的人在赢钱概率高时展现出风险厌恶，赢钱概率低时展现出风险偏好，而经历情境的人在赢钱概率高时展现出风险偏好而在赢钱概率低时展现出风险厌恶。这种风险厌恶和偏好在损失情

境出现了反转，即描述情境的人在输钱概率高时展现出了风险偏好，输钱概率低时展现出风险厌恶；而经历情境的人在输钱概率低时展现出了风险偏好而在输钱概率高时展现出了风险寻求。

The Fourfold Pattern of Risk Attitudes in Decisions from Description and it's Reversal in Decisions from Experience (Based on Hertwig et al., 2004)

	Decisions from description		Decisions from experience	
Probability	Gain domain	Loss domain	Gain domain	Loss domain
Low	32, .1[a] vs. 3, 1.0 Rare event: 32, .1 Risk seeking 48%[b]	−32, .1 vs. −3, 1.0 Rare event: −32, .1 Risk averse 36%	32, .1 vs. 3, 1.0 Rare event: 32, .1 Risk averse 20%	−32, .1 vs. −3, 1.0 Rare event: −32, .1 Risk seeking 72%
High	4, .8 vs. 3, 1.0 Rare event: 0, .2 Risk averse 36%	−4, .8 vs. −3, 1.0 Rare event: 0, .2 Risk seeking 72%	4, .8 vs. 3, 1.0 Rare event: 0, .2 Risk seeking 88%	−4, .8 vs. −3, 1.0 Rare event: 0, .2 Risk averse 44%

[a] For the sake of brevity, the alternative outcome (0 otherwise) has been omitted for all risky options. [b] Proportion of risky choices. In past studies, this proportion has been found to be greater than 50% (e.g., Tversky and Kahneman (1992)).

图 2 收录自(Wulff et al., 2018)关于描述和经历差距

其中左侧为描述情境，右侧为经历情境，研究者发现即使当期望收益相同时，被试的风险偏好会受到所处情境，风险选项以高或低概率呈现及正负效价等因素的影响。

进一步观察发现，当在收益和损失两种效价下呈现具有相同的期望收益绝对值的选项，描述和经历情境下两组被试间的差异仍具有显著不同。因此可以预测，处于描述和经历情境下的两组被试的主观效用曲线的扭曲程度也会存在差异。更进一步，从强化学习的观点出发，主观效用曲线扭曲程度的不同将会带来学习率的差异，如下**图 3** 所示。例如，当被试处于经验情境时，在收益和损失环境下产生相同的预测误差时，收益环境将比损失环境下更新的 Q 值幅度更大（由于更大的学习率），在描述情境下相反。

上述关于描述和经验情境的差异研究主要聚焦于人们的单次决策任务当中，即人们当前的决策并不会对下一次决策产生影响。然而，探究情境所引起的决策差异在序列性的不确定性价值决策任务中也具有重要意义。在这种任务中，人们往往需要不断与环境交互，通过环境给予的反馈进行学习。人们将逐渐形成关于环境结构的表征，这种表征既可以是无模型的(Model-free)，纯粹关于状态或动作价值的函数，也可以是基于模型的(Model-based)，基于状态到状态间的转移概率(Transition probability)。事实上，人们的决策往往会受到认知和计算能力的约束。因此，人们也有时会采取一种更为简单的启发式(Heuristic)策略，通过采取节省计算性的方式来进行决策(Kahneman et al., 1982)。当然，在节省计算和认知资源的同时，这也导致人们在决策过程中便会对选择的表征出现扭曲(Distortion)，从而系统性地偏离(bias)最优决策。

人们关于环境的表征并非一成不变，而会随着任务的进行而不断发生变化。在序列决策的环境中，人们会对环境结构的进行动态性的表征和适应性学习(Adaptive learning)和调节。例如，人们会通过当前环境的随机性(Stochasticity)和波动性(Volatility)来动态调节自己

的学习率，进而改变自己的行为和策略，最大化自己的目标(Piray & Daw, 2023)。人们也能够区分背后的环境模型是否发生重大改变(Nassar et al., 2010)。

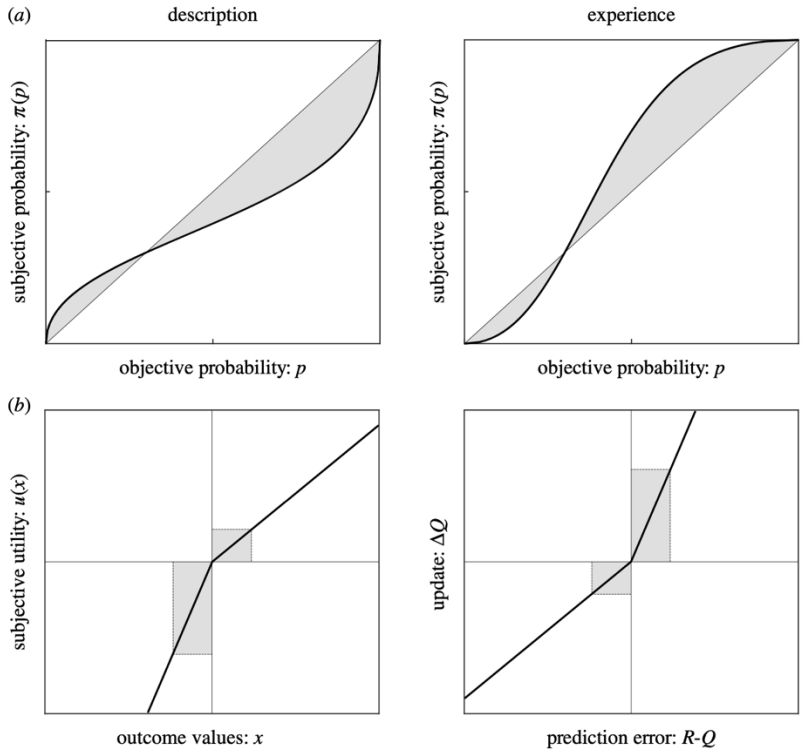

图 3 收录自(Garcia et al., 2021)关于描述和经历差距

(a)图分别展现了在描述和经历情境下被试对于主观概率的扭曲，即描述情境下的被试对于低概率高估，高概率低估，而经历情境下的被试对于低概率低估，高概率高估。(b)图展现了强化学习中对于价值更新的不对称性。在经历情境下，正负 domain 产生相同的预测误差(prediction error)时，Gain domain 更新的 Q 值要比 Loss domain 更快，描述情境下则相反。

除了对于环境的表征可能发生变化，人们的决策也可能受到之前选择和决策的影响。例如，Rouhani 等人(2020)发现，较大的预测误差可能被人类的记忆系统所强烈的编码，并且作为事件的边界中断了事件的整合；Bornsterin 等人(2017)发现，人们在复杂情境中的学习不光通过增量式，还可能通过从记忆系统中不断抽样地方式学习。

本研究主要想探究的问题在于情境在探索和利用权衡任务中的影响。探索和利用之间的权衡本身便作为一种关于如何进行信息搜索和采集的策略，将会影响智能体对于环境和收益的感知，体现为状态或动作价值函数的更新。我们认为，情境作为外部变量，能够宏观地调控被试在环境中的学习和决策行为。在本研究中，我们采用一种极简(minimalistic)的探索和利用权衡任务(Song et al., 2019)，这种极简的范式允许我们剥离掉一些无关因素，如环境的不稳定性，波动性的影响，更为纯粹地观察描述和经历环境下被试在序列决策任务下搜索或采样策略的差异性。

我们认为，在序列决策任务中，描述和经验情境的差异性体现在多个方面：首先，在经验性决策中，人们往往只会依赖有限的样本来进行选择，这可能导致对罕见事件估计的不准确。因此，环境变量可能在不同情境间对被试的选择的影响会出现差异。第二，在有关记忆，信念更新和判断研究中往往观察到，被试会较为依赖最近获得的信息来更新自己的决策，即近因效应(Recency effect)。在描述和经验情境中，由于其掌握关于环境信息的不同，描述情境下的被试可能更加相信客观的信息呈现；而经验情境下，人们往往得到的是关于事件的序列，并没有明确的概率总结，因此被试可能更依赖于之前所得到的新信息。这种信息呈现方式的差异可能会出发不同的认知算法，从而导致不同的决策行为。另外，描述情境下的被试还有可能将客观信息同经验信息结合起来进行考虑。

综上，本研究建立在前人研究的基础上，采用一种较为简化的探索和利用权衡范式，试图探究情境对于人类序列决策中的影响，并提出计算模型试图解释人类背后的认知计算模型。

第二章 研究方法

2.1 实验任务

本实验在配有 20-in ViewSonic 显示器，分辨率为 1920×1080，屏幕刷新率为 60Hz，操作系统为 Windows7 的电脑上使用火狐浏览器(Firefox)进行实验。在实验中，被试被要求想象以下一个场景**(图 4)**：想象自己是一名去外国城市旅游的游客，在每个回合（试次）中，被试在这座城市停留的天数 T 从 5-10 天不等。在每一天当中，被试都需要选择去一个餐厅吃饭。对于被试而言，除第一天外，被试每天都面临着两个选择：可以选择目前已知评分最高的餐厅吃饭(Go to the best restaurant so far)，或者去到一个随机的新餐馆(Go to a random new restaurant)。在第一天中，被试只能选择去到一家随机的新餐馆中。在被试每次点击两个选项其中之一后，在屏幕下方会出现确认按钮。被试在每个回合中都需要点击确认按钮来确定选择，以防止被试误触或不认真完成任务。在每次选择确认后，餐馆的评分将会在屏幕上呈现。此外，屏幕上呈现的信息还包括：当前回合剩余可以选择的天数 t_{left}，当前回合已知最高的评分 r^*（将用红色标记），当前回合所有餐馆评分的历史记录（试次中所有最高评分将都会被用红色进行标记），以及当前回合的得分总和。在试次结束后，被试被告知当前回合的总得分后，方可进入下一个回合中。被试在完成实验任务中不设置时间限制，其可以自己掌控完成任务的进度和快慢。

本实验采用 2（情境）×2（效价）的双因素组间设计，所有被试被随机分配至四个组别当中的一个完成实验。情境变量分为描述和经历两种情境，在描述(Description)情境中(**图 4A**)，被试能在屏幕中看到整座城市餐馆评分的分布情况，在经历(Experience)情境中(**图 4B**)，餐馆评分分布对被试并不可见，因此被试需要通过不断和环境交互来学习整座城市的餐馆的分布情况。在实验前，被试被告知整座城市餐馆的分布并不会发生改变。效价变量分为正效价和负效价环境，在正效价框架下，整座城市餐馆的评分分布遵循以 3 为均值，0.6 为标准差的截断正态分布(Truncated gaussian distribution)，餐馆的评分只从 1 分至 5 分当中抽取。在负效价框架下，整座城市餐馆的评分分布遵循以-3 为均值，0.6 为标准差的截断正态分布，餐馆评分将从-5 分至-1 分中抽取。

2.2 实验程序

在实验开始前，被试需要阅读指导语部分的内容，指导语部分均用中文进行呈现。在每种不同条件下，被试的指导语会做出一定区分。具体而言，在描述条件下，被试会被告知餐馆的真实分布的形态，并以图片的方式展示整座城市的餐馆分布，但在经历条件下，被试仅会被告知餐馆的评分将从 1 分至 5 分(或-5 分至-1 分)中抽取。负效价环境下，为更贴近真实情境，被试被告知整座城市餐馆并不符合自己的口味，但被试仍需要从中选择尽可能更高的评分以满足自己口味的需求。为测试被试是否真正理解实验任务，在指导语部

分中插入了几道测试题目。例如，在描述情境下的被试会完成以下简单的二分选择题：如果你选择随机地去到一座餐馆，哪一种你认为更有可能？A. 2.5 B. 4.5(正确答案应为 A)。在指导语部分，我们还一一对屏幕上出现的提示的含义一一作出解释。之后，我们询问了被试对于任务细节的一些理解。例如，在给定情况下，选择“去到最好的餐厅”将会得到怎样的评分？或者“如果选择去到一家随机的餐馆，是否能够得到和之前一样的评分？”等问题。只有通过测试的被试才能够进行之后的正式实验。在正式实验中，被试一共需要完成 180 个试次，每个试次包含 5-10 次选择(每种长度的试次共 30 个)。在完成实验后，我们询问了被试在实验中采用的策略。在计算被试在实验中获得的报酬中，我们随机挑取 180 个试次中的 1 个试次中的平均奖励来计算被试所获得的额外报酬情况，以激励被试更好的完成任务。在指导语部分，我们也着重强调了每个试次对于决定最终报酬的同等重要性。

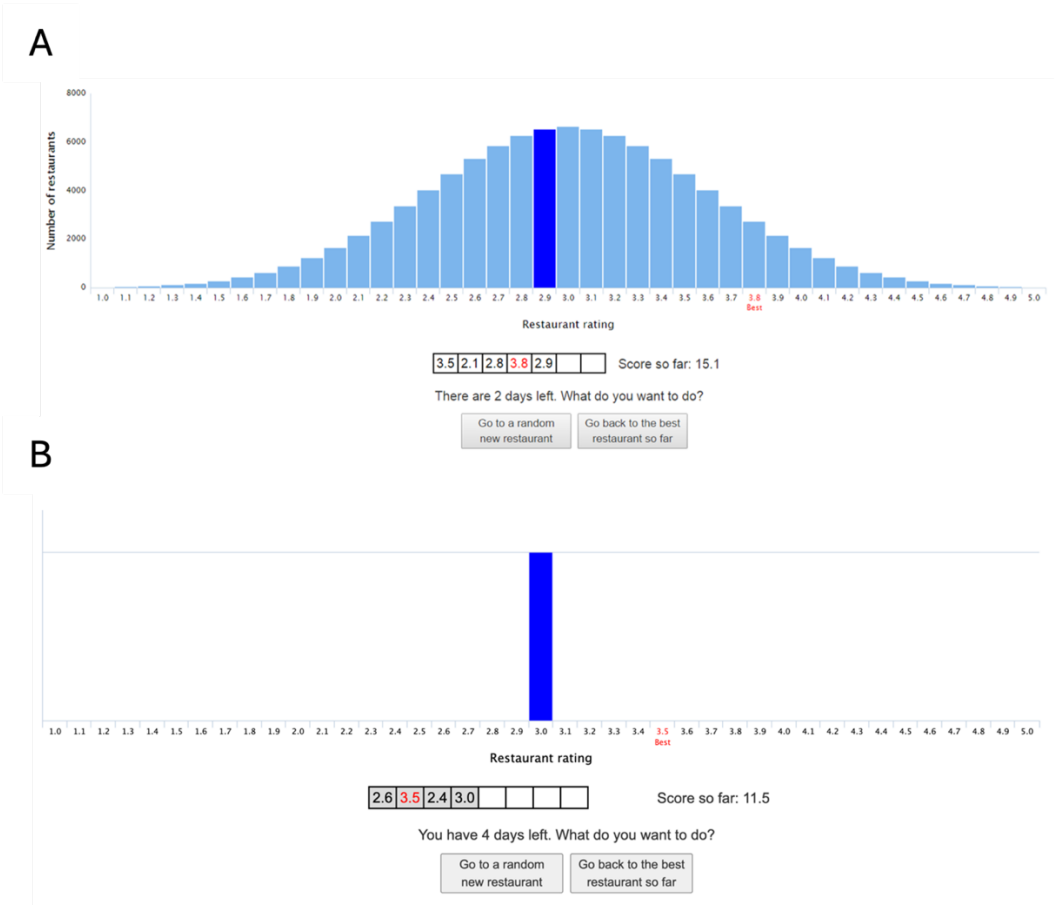

图 4 描述和经历情境下的实验范式

A:描述情境下呈现给被试的界面。在描述情境中，被试可以看到城市中每一个评分下餐馆的数量。**B：**经历情境下呈现给被试的界面。在经历情境中，被试只能知道当前所处餐馆的评分在坐标轴中出现的位置。在两个情境中，其他有关实验任务的信息均相同呈现给被试，例如当前试次选择的历史信息，当前试次的最大评分（用红色标注），试次的总得分，剩余天数等。在实验中，被试在每一次选择中可以选择去到一个评分未知的新餐馆或回到目前为止已知评分最好的餐馆。

2.3 被试

我们在实验中总共收集了 177 名被试，将被试随机分配到 4 个组别当中的其中一组，即采用组间设计，每个组别被试的数量如下**表 1** 所示。在经过解释后，每名被试均通过了指导语部分的测试内容。本项实验由北京大学心理与认知科学学院伦理委员会批准。在实验开始前，被试均在知情同意书上签字，告知本项实验中被试拥有的权力。被试的报酬由两部分组成：基础报酬为 30 元，以及额外的 0-30 元浮动报酬，由被试在任务中的表现而定（即在实验程序中的描述）。每名被试在实验中呈现的试次顺序随机化处理。

表 1 不同情境和效用条件下的被试的人数

组别	1	2	3	4
情境	描述	描述	经历	经历
效价	正	负	正	负
简称	DG	DL	EG	EL
人数	45	44	44	44

2.4 分析方法

实验任务的代码改编自 Song 等人(2019)的研究，通过 JavaScript, html 和 css 语言进行改写。数据的预处理部分使用 Python 进行。绘图和回归分析使用 R 语言和 MATLAB 进行，计算建模部分使用 Python 进行。

在进行模型拟合时，我们使用了 Python 中的 ConstrNMPy 库，通过单纯形法，拟写似然函数，通过 MLE 实现有约束的优化问题，寻找最优的参数。每名被试在每个候选模型中一共拟合 20 次，每次的初值从随机点出发，避免陷入局部最小值。

第三章 结果

3.1 局部调节：试次内情境和效价对于行为的影响

在前人的研究中，研究者已经发现被试的探索和利用行为会显著地受到当前试次的最大奖赏r^*，当前试次剩余天数t_{left}以及试次长度T的影响。这里，我们为了探究和比较在不同情境（描述/经历）和效价（正/负）条件下被试探索和利用行为的差异性，我们重复了前人研究中的多元逻辑斯蒂回归(Multinomial Logistic Regression)模型(Model1)，以检验不同变量对于被试决策的影响：

$$action \sim \beta_1 r^* + \beta_2 t_{left} + \beta_3 T$$

其中，*action* 代表被试的选择，我们将其编码为 0 或 1，其中 1 代表探索（选择随机去一家新餐馆），0 代表利用（选择去目前已知最好的餐馆）。r^*代表目前为止得到最好的餐馆评分，t_{left}代表当前试次下剩余可以决策的天数，T代表试次的总长度。**图 5** 展示了在不同最大奖赏r^*和剩余天数t_{left}下被试的探索行为的差异性。

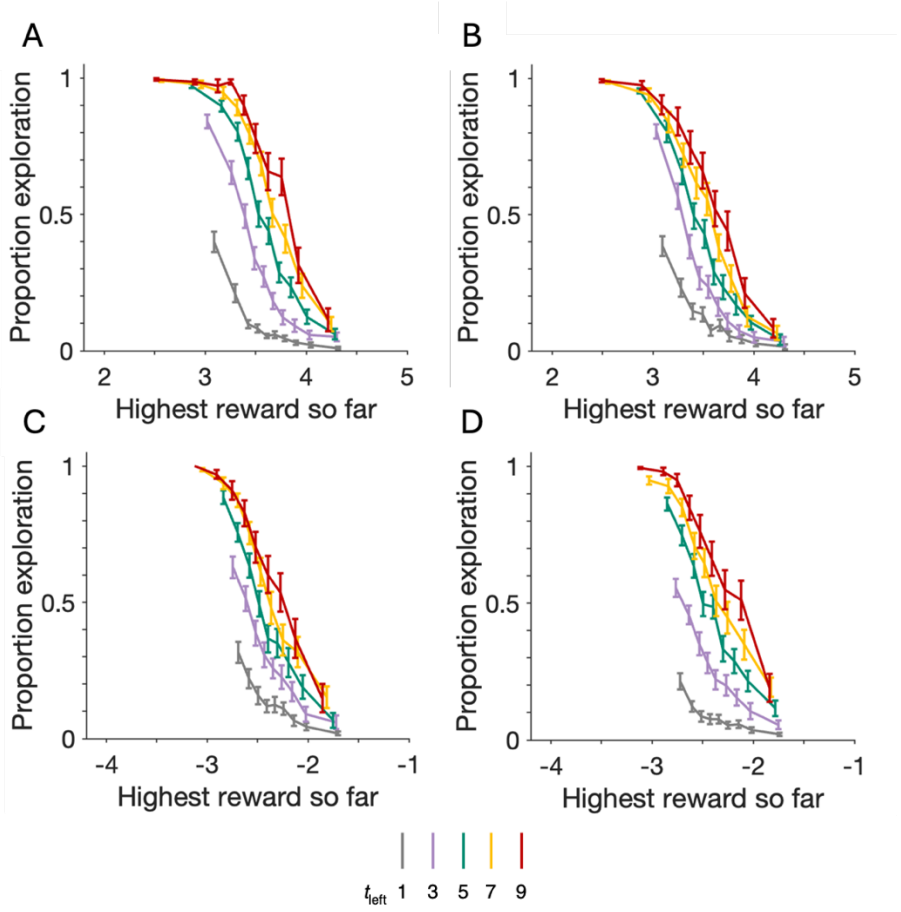

图 5 不同情境和效用下被试探索行为受到当前最大奖赏r^*和剩余天数t_{left}的影响

图 A-D 分别代表了 DG,DL,EG,EL 条件下被试的表现，不同颜色的线代表在不同剩余天数t_{left}下被试探索行为的比例

3.1.1 当前试次最大奖赏对探索行为的影响

在 Song 等人(2019)的研究中，其发现被试的行为会受到当前试次最大奖赏r^*的影响，即当前餐馆的评分r^*较大时，被试选择回到这一餐馆（利用）的比例越高，反之则会选择继续探索新的餐馆。我们在实验设置的四个条件下也均观察到了这一现象，即所有组别中最大奖赏r^* 的系数值均显著不等于 0(p<0.001)。

为了比较被试探索行为如何受到不同情境和效用条件的影响，我们将上述逻辑斯蒂回归模型中拟合的单被试的系数值进行了两两间的独立样本 t 检验，如**图 7A** 所示。我们发现，被试在经验情境(Experience Context)下受到当前试次最大奖赏r^*的影响更小，即相较描述情境（Description）下的被试展现出了更多的风险偏好(Risk averse)。具体而言，在正效价条件下，DG 和 EG 组间具有显著差异(β_{DG}=-3.607±0.296, β_{EG}=-2.832±0.245, d =-0.775, t_{79}=-2.011, p = 0.0477; Cohen's d = -0.447, 95% CI on d: -1.543 to -0.008)。在负效价情境下，DL 和 EL 组间具有显著差异(β_{DL}=-3.469±0.308, β_{EL}=-2.330±0.210, d =-1.138, $t_{72.017}$=-3.049 , p = 0.003; Cohen's d = -0.666, 95% CI on d: -1.882 to -0.394)。在相同情境下，当前试次最大奖赏对被试探索行为的影响在不同正负效价框架不具有显著差异(描述情境：β_{DG}=-3.607±0.296, β_{DL}=-3.469±0.308, d =-0.139, t_{81}=-2.011, p = 0.7457; Cohen's d = -0.071, 95% CI on d: -0.991 to -0.713；经历情境：β_{EG}=-2.832±0.245, β_{EL}=-2.330±0.210, d =-0.502, t_{79}=-1.557, p = 0.1235; Cohen's d = -0.346, 95% CI on d: -1.1432 to 0.1397)。

3.1.2 当前试次剩余决策天数对探索行为的影响

被试的行为除会受到当前试次最大奖赏r^*外，还会强烈受到试次剩余天数 t_{left} 的调节，如上**图 5** 所示。图中不同颜色的曲线代表了在不同剩余天数 t_{left} 水平下被试探索行为比例的差异。在同一餐馆评分下，若当前试次剩余天数 t_{left} 较少，被试则更有可能利用当前已知的最大餐馆的评分，而剩余天数 t_{left} 较多时，被试则更愿意去探索新的餐馆。这一结果也与前人研究的结果相符。

尽管我们假设被试的决策还会受到当前试次的剩余选择天数的影响，但我们在逻辑斯蒂回归模型中并没有发现不同情境和效用框架下被试决策受到剩余天数 t_{left} 的差异。我们采用两两之间的独立样本 t 检验，如**图 7B** 所示。在正效价条件下，DG 和 EG 组间不具有显著差异(β_{DG}=0.745±0.276, β_{EG}=0.852±0.332, d =-0.107, t_{78}=-1.5706, p = 0.1203; Cohen's d = -0.351, 95% CI on d: -0.243 to -0.0287)。在负效价情境下，DL 和 EL 组间不具有显著差异(β_{DL}=0.802±0.232, β_{EL}=0.896±0.288, d =-0.095, t_{74}=-1.575, , p = 0.1196; Cohen's d = -0.361, 95% CI on d: -0.214 to -0.025)。此外，在相同情境下，当前试次剩余选择天数对被试探索行为的影响在不同正负效价框架不具有显著差异(描述情境：β_{DG}=0.745±0.276, β_{DL}=0.802±0.232, d =-0.058, t_{76}=-0.995, p = 0.323; Cohen's d = -0.225, 95% CI on d: -0.173 to

0.058；经历情境：β_{EG}=0.852±0.332, β_{EL}=0.896±0.288, d =-0.045, t_{76}=-0.636, p = 0.527; Cohen's d = -0.144, 95% CI on d: -0.186 to 0.096。

3.1.3 当前试次总天数对探索行为的影响

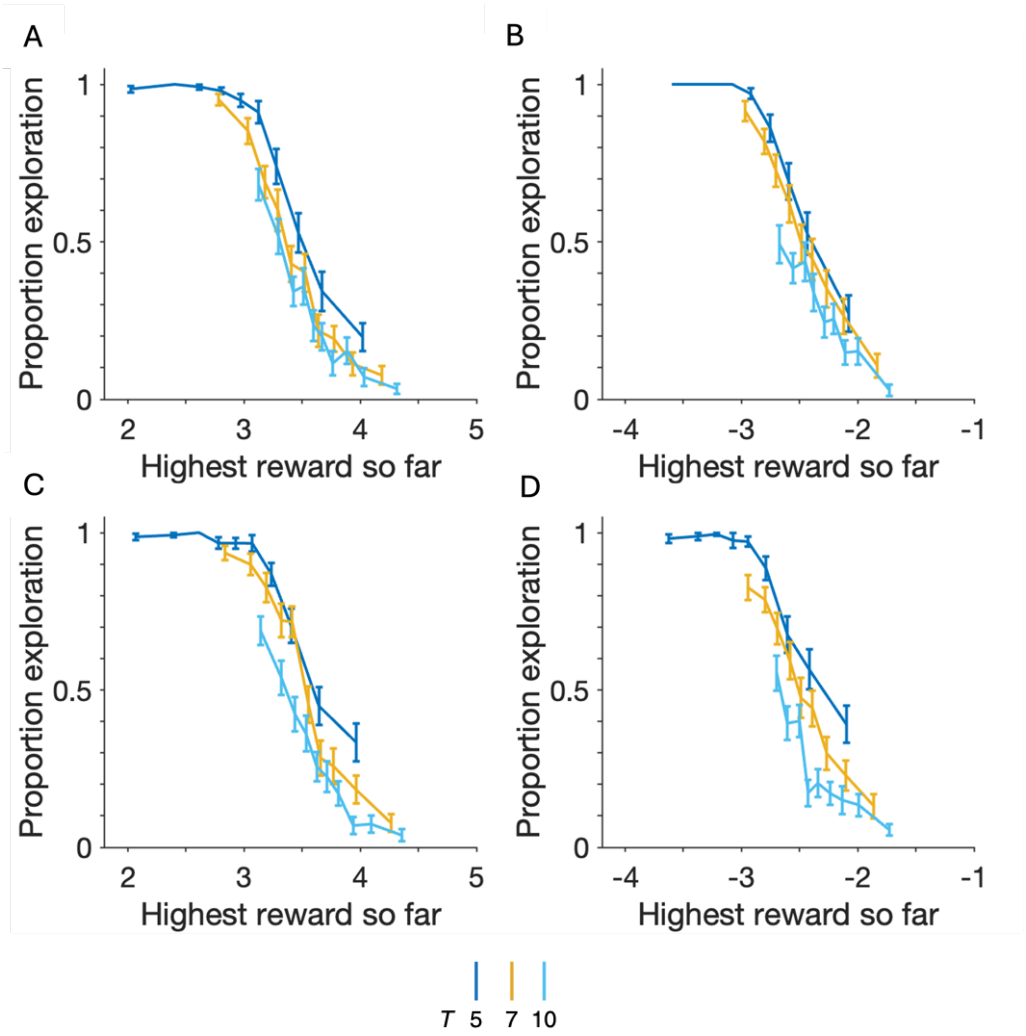

图 6 不同情境和效用下被试探索行为受到当前试次长度 *T* 的影响

图 A-D 分别代表了 DG,DL,EG,EL 条件下被试的表现，图中不同颜色代表了在不同试次长度 *T* 下被试探索行为的比例

被试的行为除受到当前试次最大奖赏r^*和剩余天数 t_{left} 的影响外，还会受到试次长度 *T* 的影响。**图 6** 不同颜色曲线代表了在不同天数下试次长度下被试探索的比例。在给定的剩余天数 t_{left} 下，被试在试次更短的试次探索的比例要比更长的试次更高。同样在多元逻辑斯蒂回归分析中，四个组别关于试次总长度 *T* 的系数值均显著不等于 0(p<0.001)。

我们进一步探究了试次长度 *T* 在不同情境和效价条件下的差异性，如下**图 7C** 所示。我们发现，被试在经验情境并处于负效价的框架下，其探索和利用行为受到试次长度 *T* 的

影响更大。具体而言，在负效价框架下，经验情境和描述情境间的差异呈现边缘显著（β_{DL}=-0.389±0.188, β_{EL}=-0.471±0.209, d =0.081, t_{78}=1.802, p = 0.075; Cohen's d = 0.408, 95% CI on d: -0.0085 to 0.1710）。同样在经验情境下，正效价和负效价条件下的差异也呈现边缘显著（β_{EG}=-0.386±0.211, β_{EL}=-0.471±0.209, d =0.085, t_{77}=1.802, p = 0.076; Cohen's d = 0.404, 95% CI on d: -0.0092 to 0.1796）

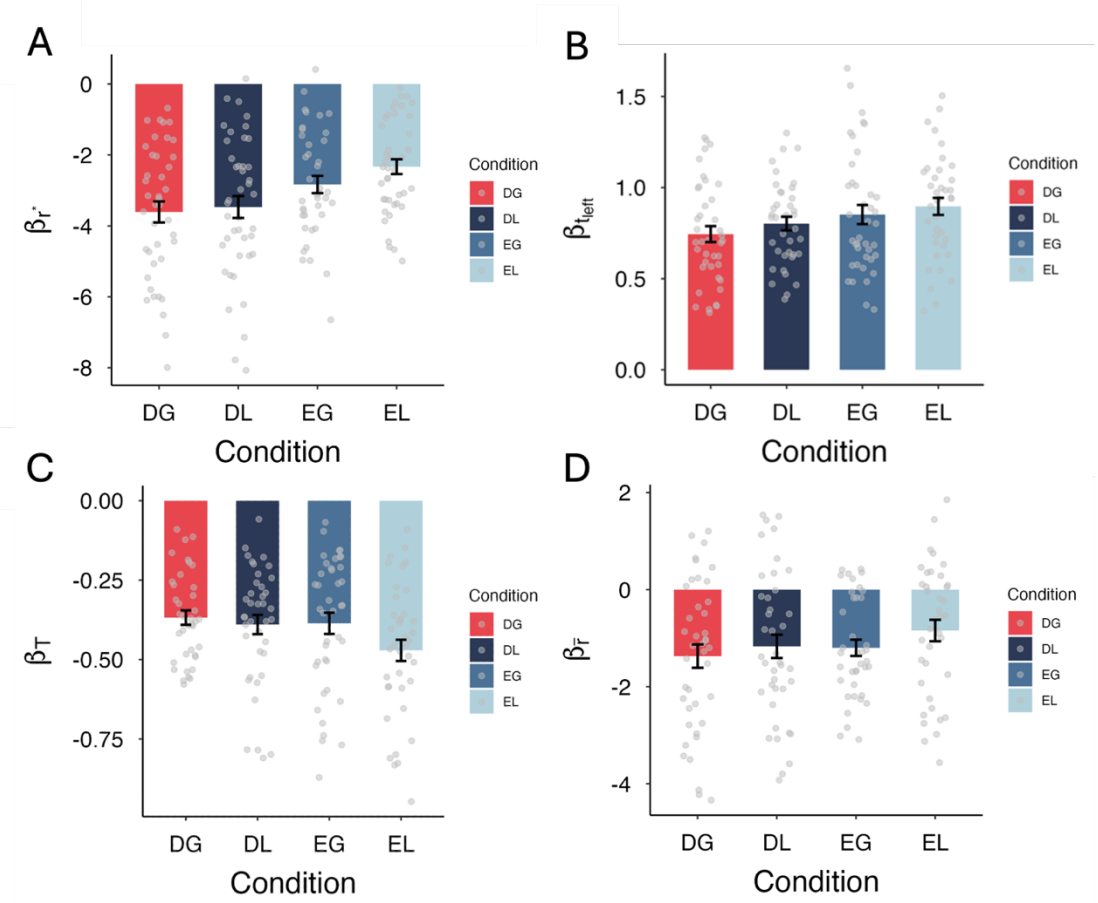

图 7 不同情境和效用下条件下回归模型当前试次最大奖赏r^*系数(A)，当前试次剩余天数t_{left}(B)，试次长度 T(C)和当前试次平均奖赏$\bar{r}$(D)的差异性

3.1.4 当前试次均值变化对探索行为的影响

我们希望在原实验的基础上发现更多可能影响被试探索和利用行为的因素及情境在其中的调节作用。我们考虑以下逻辑斯蒂回归模型(Model2)，探究被试的探索和利用行为是否会受到试次均值变化的影响，即

$$action \sim \beta_1 r^* + \beta_2 t_{left} + \beta_3 T + \beta_4 \bar{r}$$

其中$\bar{r}$代表当前试次所有餐馆的平均评分。

图 7D 展示了四种不同条件下通过回归模型拟合的$\bar{r}$的参数结果。我们发现，参数对于被试探索和利用行为的影响是显著的(DG, DL 和 EG：p<0.001; EL: p<0.01)，即当前试次奖

赏的均值更大时，被试倾向于选择利用当前已知最好的餐馆。然而，相较于当前试次的最大奖赏r^*而言，一部分被试拟合的参数值出现了正性的结果，且这种影响在单被试的层面达到显著。另外值得注意的是，当前试次的最大奖赏r^*和平均奖赏$\bar{r}$之间并不存在非常明显的共线性问题(*VIF<10*)。例如，当被试在已有一个较好的奖赏，继续采取探索行为而获得一个较差的餐馆评分后，此时平均奖赏$\bar{r}$相较于之前会有下降，而最大奖赏r^*则不会发生变化。因此，我们猜测出现上述这种平均奖赏$\bar{r}$对于被试探索行为的正性影响的原因可能是源于被试的非理性行为，即当已有较好的奖赏后仍然选择继续探索。

我们试图继续探索这种影响的根源。此处我们定义新的心理量为差距(*Gap)*，指被试在当前试次的第 *i* 次选择时所获得的奖赏相较于当前最好奖赏r^*之间的差异。例如，如果被试在 *i* 次选择探索动作，即选择去到一个随机的新餐馆，而获得的餐馆评分为 3.3，而当前最好的评分为 3.6 时，则此时的 *Gap* 值为-0.3。我们认为 *gap* 会对被试进行第 *i+1* 次选择时造成影响。此外，可以注意 *Gap* 值会存在正负之间的差异，因此，当所获得的奖赏大于当前最好奖赏而获得正的 *Gap* 值相较所获得奖赏小于当前最好奖赏而获得负的 *Gap* 值时，这对被试的探索行为的影响可能存在差异性。下**图 8** 分别展示了在 DG 组和 DL 组下 *Gap* 的整体分布情况，及每名被试和群体水平下不同 *Gap* 水平下探索的比例。如**图 8B** 所示，我们发现描述和经历组的被试在正负 *Gap* 的绝对值相同的情况下均出现了对于探索行为的不对称性，例如，在 *Gap* 的绝对值均为 1 时，正 *Gap* 相比负 *Gap* 探索的比例更低。

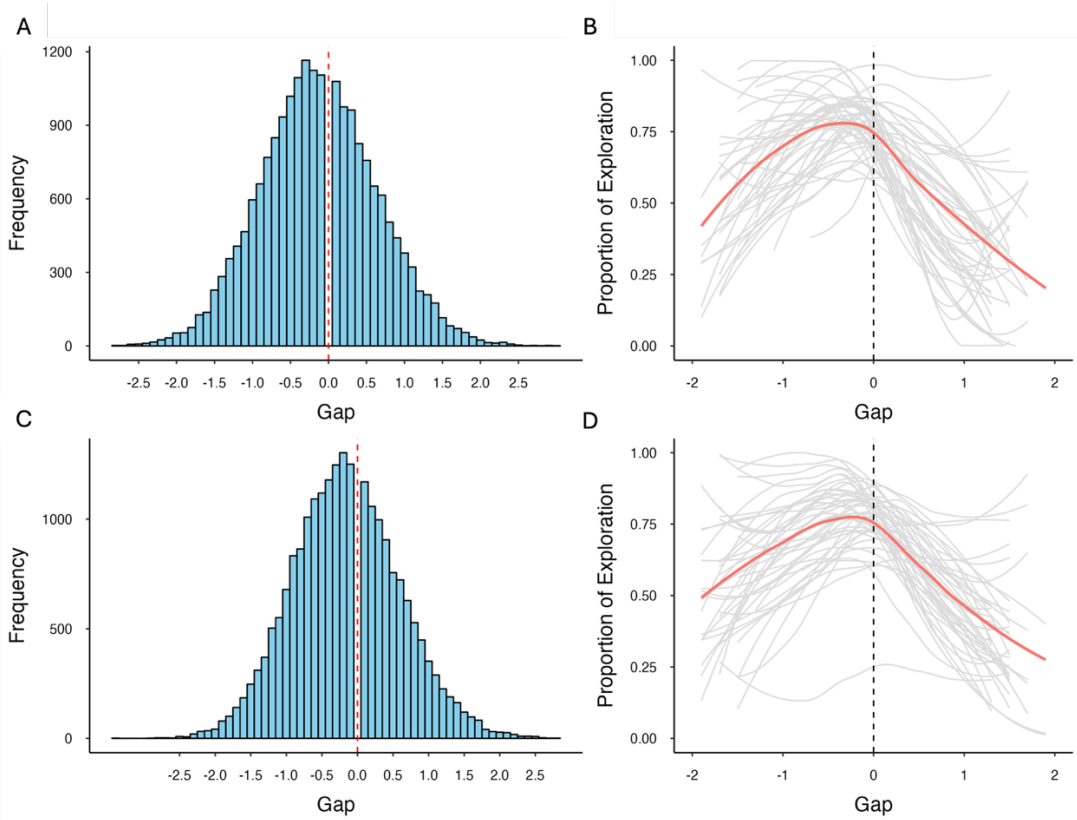

图 8 *Gap* 的整体分布和不同 *Gap* 水平下探索行为的比例(以 DG, EG 组为例)

A 图和 C 图分别展示了 DG 组和 EG 组 *Gap* 整体上呈现正态分布的形状，全距大致为 5。B 图和 D 图分别展示了 DG 组和 EG 组在不同 *Gap* 水平下每名被试探索行为的比例(灰色)和群体水平上被试探索行为的比例(橘色曲线)。图中可以看出描述和经历情境下的被试在 *Gap* 为正或负时存在不对称性。

我们希望进一步观测在描述情境下被试的探索行为是如何受到 *gap* 值的影响，我们绘制了在不同最大奖赏 r^*和 *gap* 水平下被试的探索行为，如下**图 9** 所示。**图 9A** 中不同颜色的线条展示了在 *gap* 水平为正数时，探索比例随最大奖赏 r^*的变化趋势。我们发现，不同 *Gap* 值下被试探索行为的斜率有一定区别。整体上，当 *Gap* 为正时，被试在更小的 *Gap* 值下相较于更大的 *Gap* 值下更倾向于探索；相反，当 *Gap* 为负时，被试在更大的 *Gap* 值下相较于更小的 *Gap* 值下更倾向于探索。总结而言，当 *Gap* 的绝对值更小时，被试更倾向于探索环境，而 *Gap* 的绝对值较大时，被试则相对保守，利用当前已知最好的餐馆。此外，我们仍然发现了**图 8** 中论述的不对称性的存在，即在 *Gap* 值的绝对值相同时，负 *Gap* 下被试探索的倾向越高。值得注意的是，这种不对称性仅会在最大奖赏 r^*较大时才会存在，**图 9B** 和**图 9D** 能够较为观测到这一点。

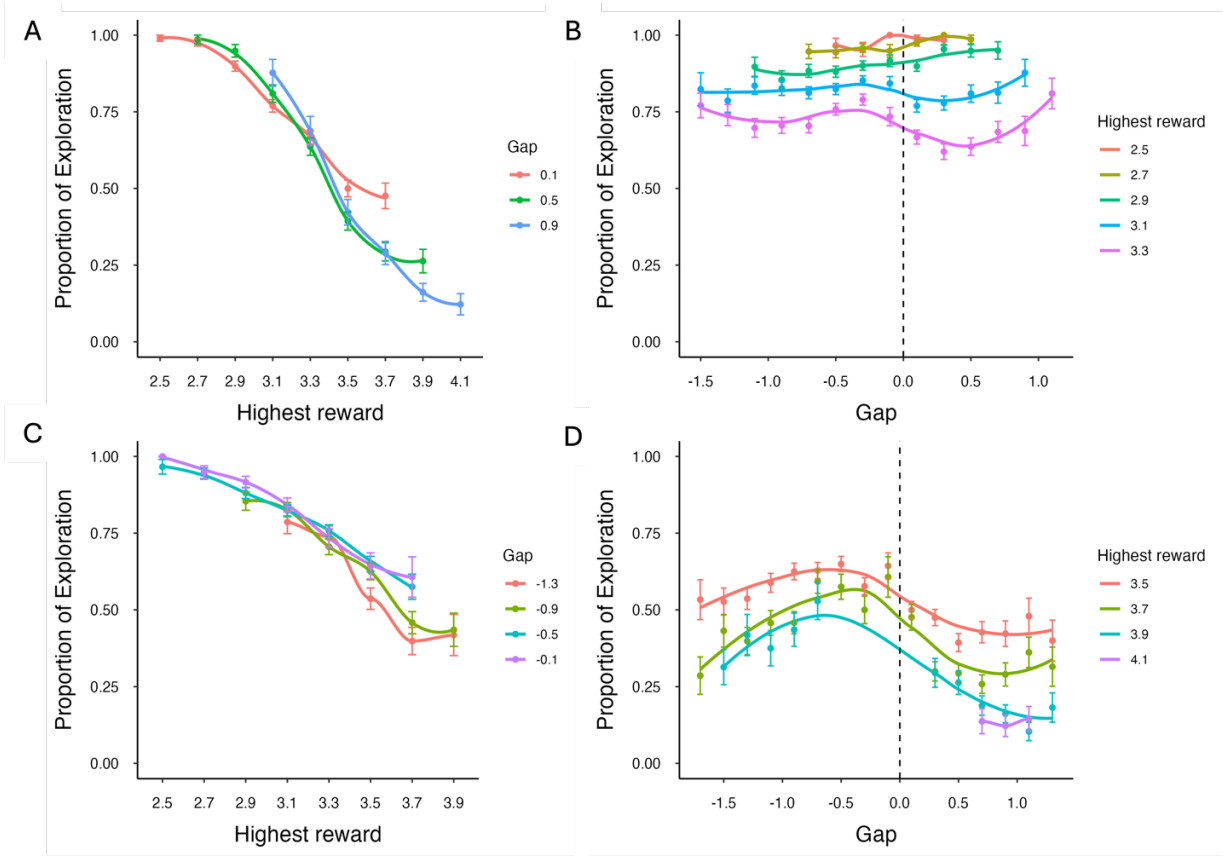

图 9 DG 组中不同 *Gap* 水平和当前试次最大奖赏下被试的探索行为的比例

A: 在 *Gap* 为正时被试在不同最高评分下探索的比例，可以看出当评分较小时更大的 *Gap* 使被试的探索倾向更大，而较大评分时，更大的 *Gap* 使被试探索的倾向更小。B：评分较小时不同 *Gap* 下

被试探索的比例，基本上并没有出现明显的不对称性。C：在 *Gap* 为负时被试在不同最高评分下探索的比例，相较 *Gap* 为正时整体的探索倾向更高。D：评分较大时不同 *Gap* 下被试探索的比例，可以看出一定不对称性，即相同绝对值下，负 *Gap* 值相较正 *Gap* 值其探索倾向更大。

根据以上对原始数据的观察，我们使用以下逻辑斯蒂回归模型(Model3)以检验 *Gap* 对于被试行为影响的效应：

$$action \sim \beta_1 r^* + \beta_2 t_{left} + \beta_3 T + \beta_4 \bar{r} + \beta_5 gap + \beta_6 neggap + \beta_7 gap \times neggap$$

其中，*gap* 为差距项，存在正负；*neggap* 编码为 1 当且仅当 *gap* 值为负，否则为 0；*gap*×*neggap* 为二者的交互项。

我们对上述逻辑斯蒂回归模型所拟合系数值进行统计分析，进行了单样本 *t* 检验，拟合的系数值及显著性水平见下**图 10**。我们发现，在我们所设置的四种条件下，*gap* 在群体水平对于探索行为的预测结果均显著(拟合系数值及具体 *p* 值见**附录 B 表 2 和表 3**)。此外，*neggap* 项在群体水平也均显著(p<0.001)，这也一定程度表明了正负 *gap* 之间的确存在不对称性。对拟合的系数值进行两两比较，我们发现在负效价条件下被试群体水平的系数值要略大于在正效用条件下，无论在经历条件下(t_{86}=1.859, p = 0.066, Cohen's d = 0.396)或是描述情境下，尽管这一差异并没有达到显著。这表明负效价环境下被试可能对 *gap* 更为敏感，从而调整自己的策略和行为。此外，情境可能还通过效价来间接调节被试探索的行为。具体而言，我们描述情境中正负效价条件下 *gap* 对于探索利用行为的贡献程度的差距要小于经历条件下二者之间的差距。推测原因，这可能由于描述情境下的被试会根据呈现的分布信息以及当前最大奖赏和目前奖励之间的差距来调整自己的行为，导致在正负效价环境下差距并不大，而经历情境下的被试由于并未呈现分布信息，正效价环境下的被试更具有风险偏好，而负效价环境下的被试则对与最佳评分之间的差距更为敏感。我们将 *gap* 项到模型后，回归模型拟合的 AIC 相较于原模型也具有了一定的改善（**附录 B 表 4**）。

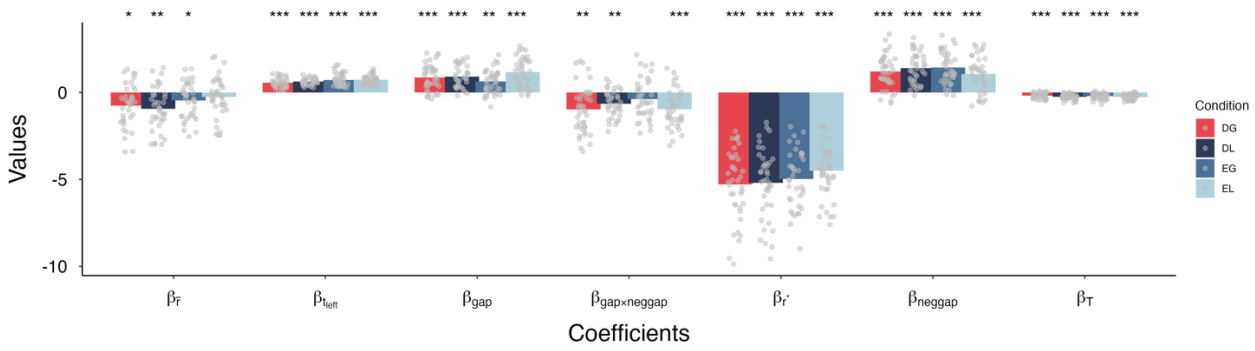

图 10 不同情境和效价条件下加入 *Gap* 项的回归模型系数

（注：*p<0.001, **p<0.01, *p<0.05）**

系数从左至右依次为平均奖赏$\bar{r}$, 剩余选择天数 t_{left}, 差距项 *gap*, 交互项 *gap*neggap*，最大评分 r^*，负差距增益 *neggap* 和试次长度 *T*

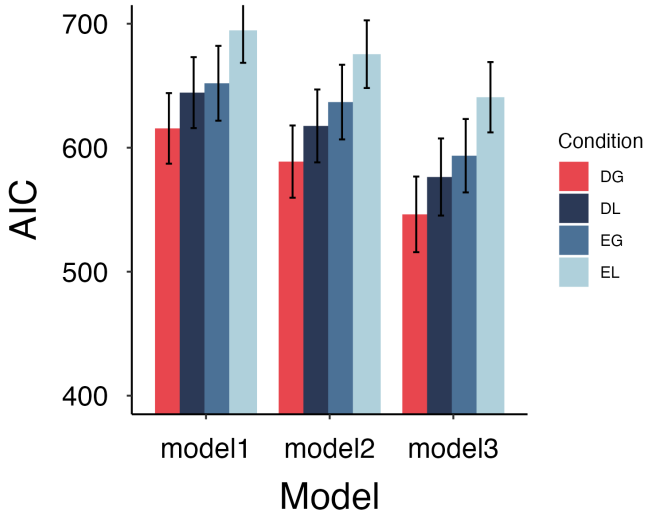

图 11 不同情境和效价条件下加入 *Gap* 项的回归模型相较于原模型拟合优度比较

Model1 为仅有最大奖赏r^*，剩余天数t_{left}和试次长度 T 的逻辑斯蒂回归模型，Model2 为 Model1 基础上加入试次平均奖赏$\bar{r}$的回归模型，Model3 为在 Model2 基础上加入 *gap* 项后的回归模型

图中可以看出 Model3 的表现在四种条件下均最佳

3.2 全局调节：试次间环境感知的动态调节

在上一章节中，我们主要讨论了在较短的时空范围内（试次内），被试如何根据当前环境的好坏即对于环境变量的感知来调节自己的行为和决策，我们称其为局部调节(Local adjustment)。正如前言所指出的，人们或许还存在更为长期的对于环境的适应和动态调节，我们将其称为全局调节(Global adjustment)。本章我们主要分析试次间被试是如何动态调节自己的行为和决策。

3.2.1 基于环境均值的动态调节

根据强化学习理论，人们可以通过试错(Trial-error)的方式进行奖赏学习。具体而言，被试会根据与环境的交互来计算每个选项背后的价值，通过一定的学习率来更新自己对于选项价值的表征，在之后的决策过程中，被试会比较不同选项带来的收益，从而作出优化的决策。在我们的研究中，我们认为被试也存在试次间类似的动态性调整。具体而言，当被试感知最近环境的收益整体变好时，会相对提高探索环境的比例，愿意去探索新的餐馆；而当整体收益不高时，则会采取相对更为保守的策略。

我们首先尝试在试次水平上（区别于之前每个选择水平上）进行相关分析。我们发现，被试在一个试次平均探索的次数与当前试次平均奖赏呈现显著的负相关关系(**图 12A**)，即当前试次获得的平均奖赏越大，则被试越倾向于利用，反之则继续探索，且探索行为比例相较上一个试次的增减和平均奖赏相较于上一个试次的增减也呈现显著的负相关(**图 12B**)。

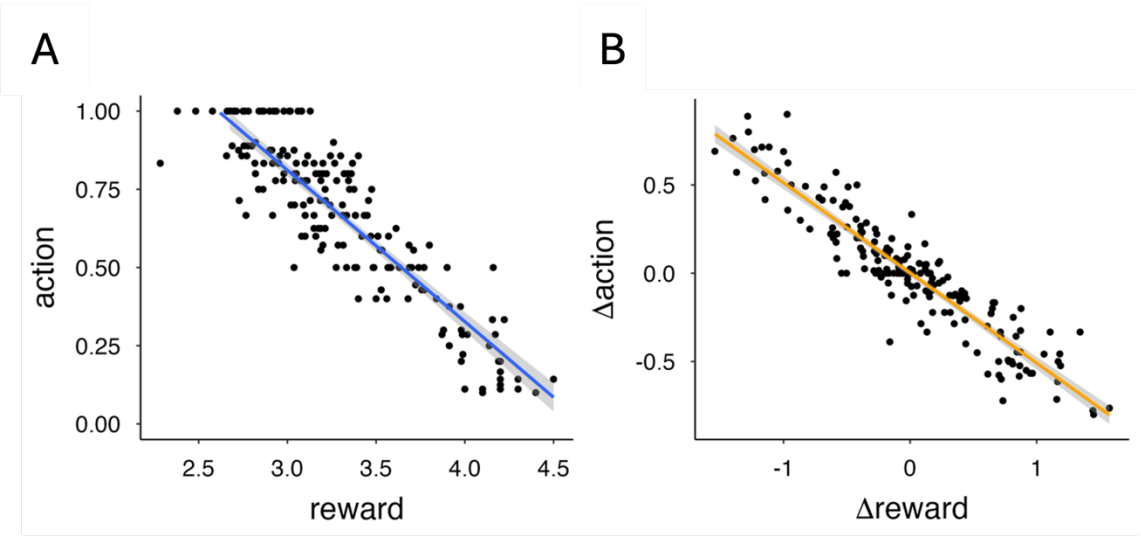

图 12 被试平均奖赏和平均探索比例之间的关系

以 EG 组的 23 号被试为例。A:当试次的平均奖赏更高时，被试探索的平均比例降低；B:当被试得到的平均奖赏相较上一试次更高时，被试对应探索的平均比例相较上一试次也会降低

有了试次水平平均探索程度和平均奖赏$\bar{r}$之间关系的证据，我们希望进一步能够观察到被试在试次间动态调节的行为。具体而言，我们假设当被试当前处于试次 *t* 时，被试会比较试次 *t-1* 和试次 *t-2* 所获得奖赏的均值差异，这里我们计算时仅选取被试采取探索动作时(选择去一家评分未知的新餐馆)获得的评分的均值，即

$$\Delta reward_{t-1}(a=1) = reward_{t-1}(\mathrm{a}=1) - reward_{t-2}(a=1)$$

同样，我们可以计算试次 *t* 和试次 *t-1* 间采取探索动作的比例差异，即

$$\Delta action_t = action_t - action_{t-1}$$

通过相关性分析，我们发现$\Delta action_t$和$\Delta reward_{t-1}(a=1)$间存在显著的正相关关系(**图 13B**)，即当被试若在 *t-1* 试次通过探索得到的评分均值更高，则被试倾向于在 *t* 试次更多的采取探索动作，反之则会采取更加保守的策略。这种效应在四种不同情境和效价条件下均存在(**图 13C**)，r_{DG}= 0.416 ± 0.092，r_{DL}= 0.411 ± 0.086，r_{EG}= 0.411 ± 0.073，r_{EL}= 0.378 ± 0.081。

为避免有其他混淆变量的干扰和随机效应，我们又采用如下多元线性回归模型(Multinominal Linear Regression)检验以二者之间的关联程度(Model4)：

$$\Delta action_t = \beta_1\, \Delta reward_{t-1} + \beta_2\, \Delta reward_{t-1}(a=1) + \beta_3 \Delta r^*_{t-1} + \beta_4 T + \beta_5 r^*$$

其中，$\Delta reward_{t-1}$为被试在 *t* 和 *t-1* 试次所获得奖励的平均值的差值，$\Delta reward_{t-1}(a=1)$为被试在 *t* 和 *t-1* 试次仅通过探索动作所获得奖励的平均值的差值，Δr^*_{t-1}为被试在 *t-1* 和 *t-2* 试次所获得最大奖励的差值，T为试次 t 的长度，r^*为试次 t 的最大奖赏。经检验，$\Delta reward_{t-1}(a=1)$的系数值仍显著(p<0.001)，即表明其能够显著地预测$\Delta action_t$的变化，而$\Delta reward_{t-1}$并不具有这种效应(四组最小值 p>0.241)。该多元线性回归的详细结果参见**附录 B 表 5** 和**表 6**。

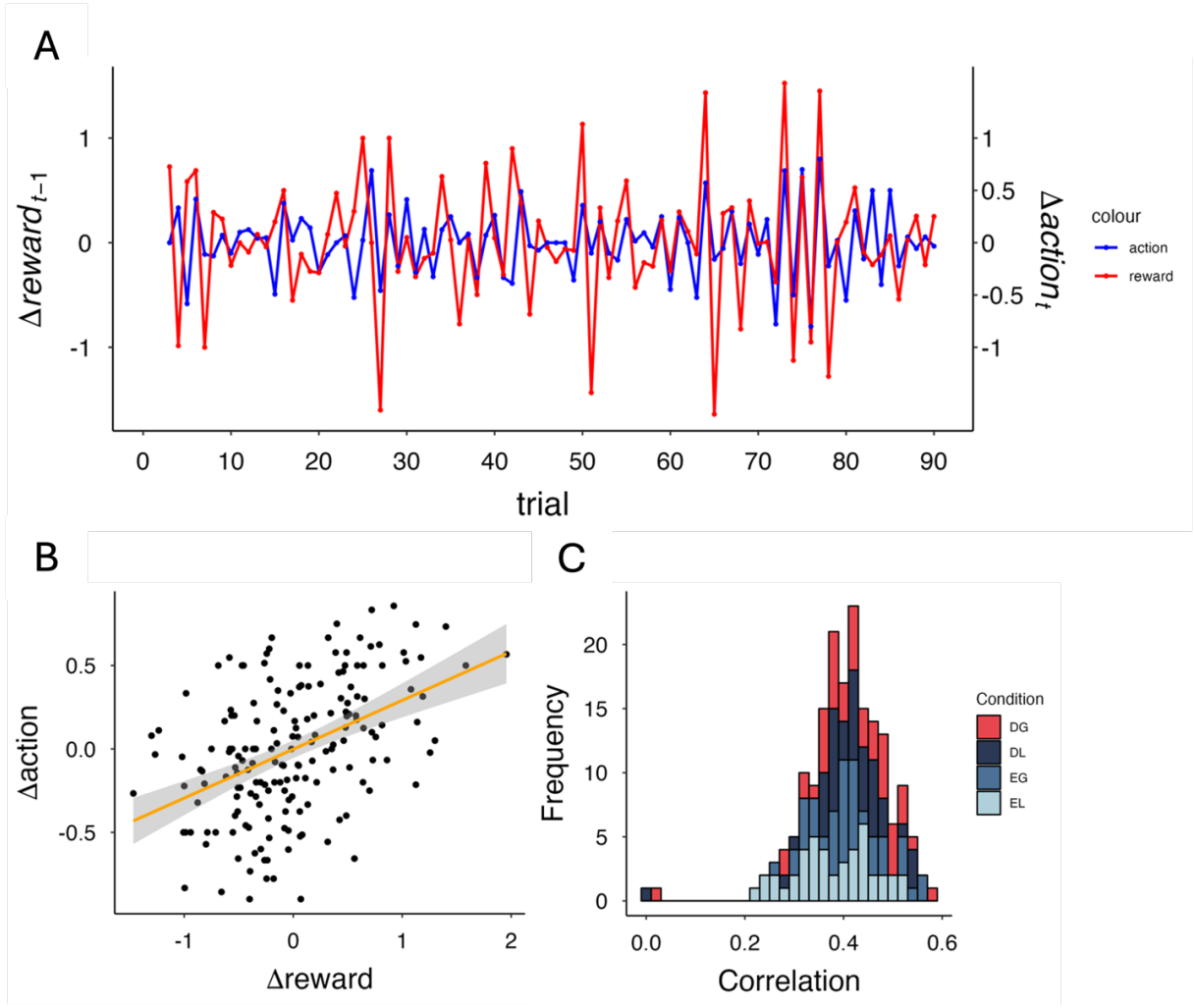

图 13 被试上一试次的奖赏变化对被试下一试次探索行为的影响

A：上一试次奖赏的差值与当前试次探索动作比例的差值之间的同步变化。B：上一试次奖赏的差值与当前试次探索动作比例的差值之间的相关。C：群体水平相关值的分布。四种条件下的被试群体水平均存在二者间的相关性。（图 13A 和 B 均以 EG 组 23 号被试为例）

我们进一步希望探究，是否更近的试次相较于更远的试次对其下一次试次探索行为的比例影响更大，即是否存在近因效应。综上，我们采用如下的多元线性回归模型(Model5)以检验近因效应是否存在：

$$\Delta action_t = \beta_1\, \Delta reward_{t-1}(a=1) + \beta_2\, \Delta reward_{t-2}(a=1) + \beta_3\, \Delta reward_{t-3}(a=1) + \beta_4\, \Delta reward_{t-4}(a=1) + \beta_5\, \Delta reward_{t-5}(a=1) + \beta_6 r^* + \beta_7 T + \beta_8 \Delta r^*_{t-1}$$

如下**图 14** 所示，我们发现前 5 个试次相较上一个试次均值的增减对于被试在下一个试次探索行为的增减的效应均显著(p<0.001)，此外，更近的试次所拟合的系数值的效应要比更远的试次的对于被试行为的影响越大。因此，被试对于选择的调节的确会受到之前被试所得到奖赏的影响，且这种影响在随时间而变小。

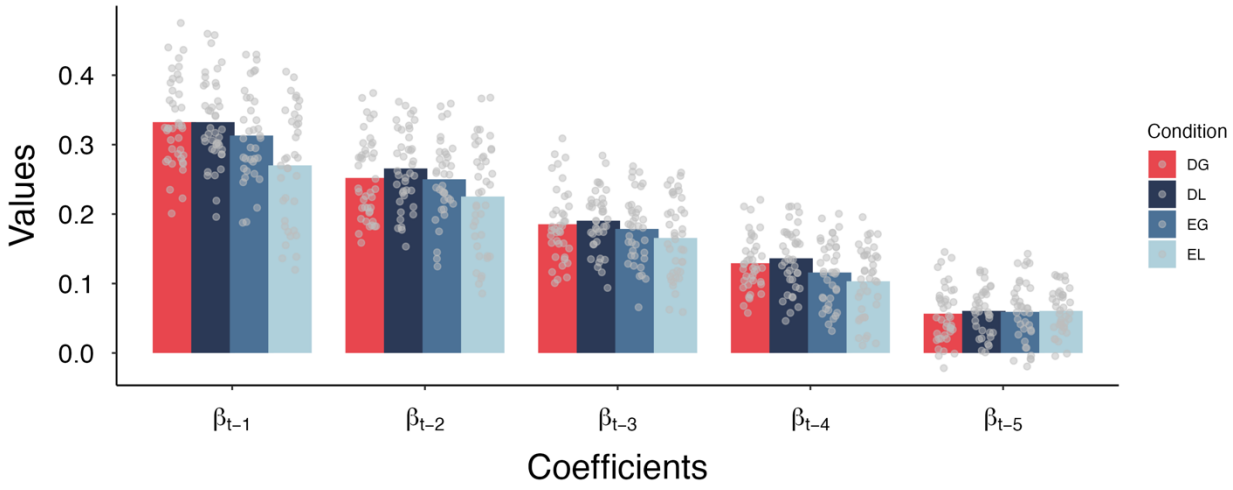

图 14 近因效应：更近的试次较更远试次对被试探索行为的影响更大

3.2.2 基于环境波动性(Stochasticity)的动态调节

在上一章节中，我们发现被试会根据前一试次奖励的增减来调整其探索动作的多少，即当整体奖赏变好时，被试更愿意探索新的餐馆；当奖赏变低时，被试对应会更保守的选择利用当前已知最好的餐馆。

除此之外，我们认为环境的随机性也是被试在进行决策时一重要考量因素。例如，考虑两个均值相同，但标准差不同的环境，我们认为被试在随机性较小的环境中可能会更加倾向于探索环境试图获取更大的奖赏。因此，我们认为被试还会根据当前试次奖励的随机性来调整探索动作的比例。例如，当在某一试次中，被试获得的奖赏的标准差较大时，被试探索的比例会下降；相反，如果获得的奖赏较为稳定时，被试则更多会探索环境。

为检验环境的随机性对被试探索行为的影响，我们进行了如下的多元回归分析：

$$action = \beta_1 reward + \beta_2 stochasticity + \beta_3 r^* + \beta_4 T$$

其中，*reward* 表示当前试次采取探索动作时获取的平均餐馆评分，*stochasticity* 表示当前试次采取探索动作时获取的平均餐馆评分的标准差。如下**图 15C** 所示，我们发现 β_1 对试次探索动作比例的效应显著(β_1 = -0.339, p<0.001)。当前试次的平均奖赏较高时，被试探索的比例较少，反之则更高。此外，当餐馆评分的平均值相同时，当前试次餐馆评分的随机性越大，则被试采取探索新的餐馆的动作会显著下降(β_2 = -0.267, p<0.001)。

据上，我们希望观察到环境随机性的效应是否在试次间仍然存在，因此，我们在模型 4 的基础上新加入了随机项$\Delta stochasticity_{t-1}$，表示试次 *t-1* 相较于 t-2 随机性的变化程度：

$$\Delta action_t = \beta_1 \Delta reward_{t-1} + \beta_2 \Delta reward_{t-1}(a = 1) + \beta_3 \Delta r^*_{t-1} + \beta_4 T + \beta_5 r^* + \beta_6 \Delta stochasticity_{t-1}$$

我们发现$\Delta stochasticity_{t-1}$的系数在四个组均显著($p$<0.001)，且系数值均为正，表明当环境变得更加随机时，被试倾向于在当前试次更多地探索环境，否则利用当前最好的奖赏。

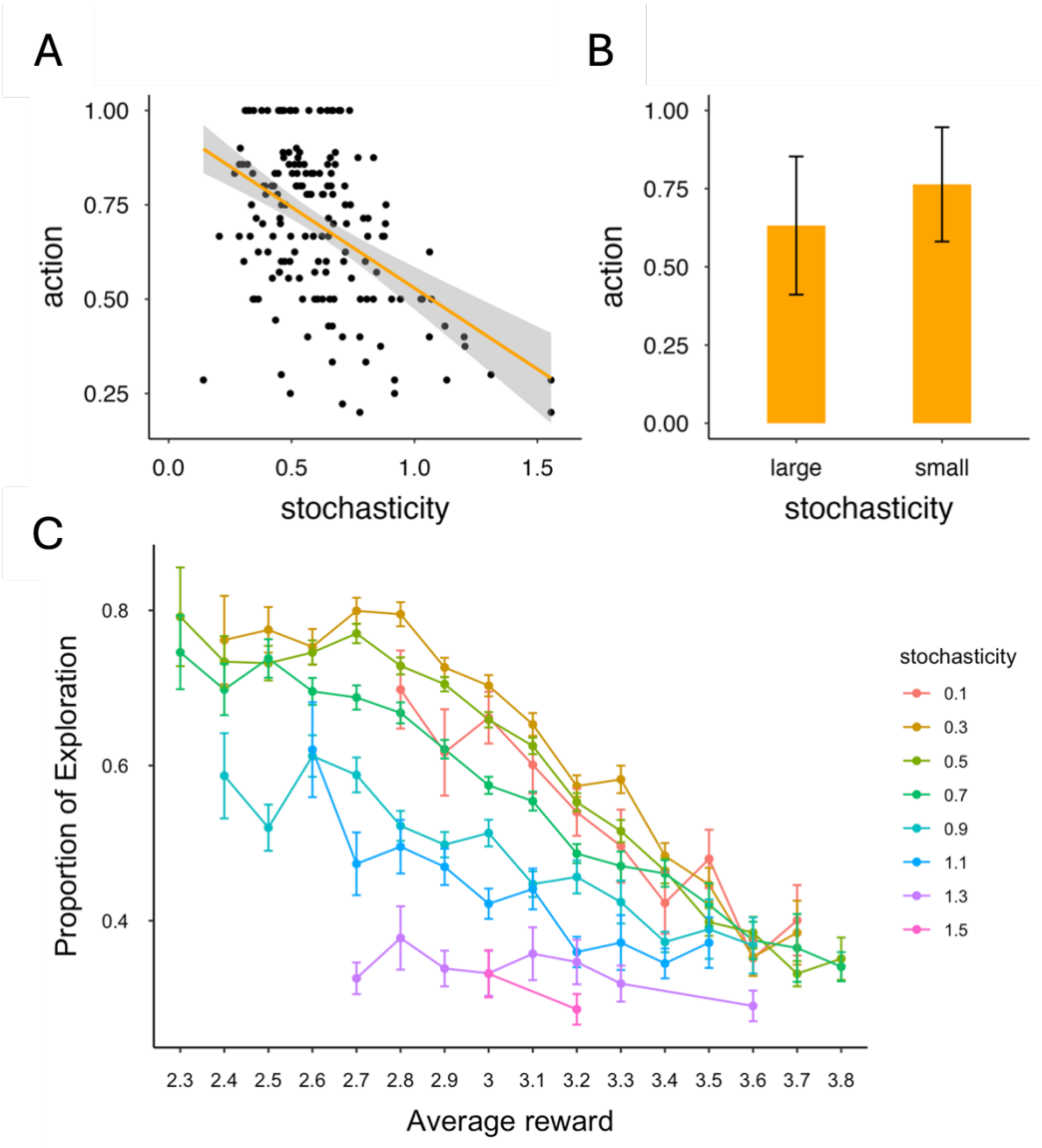

图 15 环境的随机性对于被试探索行为的影响

A：被试探索的比例和当前环境随机性的相关性。B：环境的随机性较大时相较于随机性较小时被试在试次内探索的比例会更低。C：群体水平上被试探索比例与试次平均奖赏和随机性之间的关系。当相同的平均奖赏水平下，更大的随机性使被试更少的探索环境。

第四章 计算建模

我们在上一章节中通过从原始数据样式的观察和对数据进行无模型的回归分析，我们找到了一系列有关被试如何从局部和全局水平上基于对环境的感知从而动态和适应性地调整自己的行为。基于以上分析，本章希望通过计算建模的角度来对本研究下的探索和利用任务进行建模，以探究人类如何实现动态的环境更新和决策。

在 Song 等人(2019)的研究中，其已经发现人们会系统性地偏离该环境下最优的决策策略（证明见**附录 A**）。相反，人们会采用一种更为简单的启发式策略进行决策，研究者将其称之为 Prop-V risk 模型。该模型认为，人们在每一个试次开始前会预先设定好阈值，且阈值随着剩余选择天数的减小而降低，如果被试得到的最优奖赏大于阈值，则有更大的概率选择利用，否则选择探索。此外，阈值上下还存在一定的浮动空间。

4.1 模型 1:Prop-V risk Model

在每个试次中，被试会设定一个阈值 θ，阈值会随着t_{left}/T的值减小而减小，即

$$\theta = k\frac{t_{left}}{T} + b$$

被试的决策遵循$sigmoid$函数，即

$$f(x) = \frac{1}{1 + e^{-\beta x}}$$

被试存在决策阈限，当最大奖赏超过阈值时选择探索，否则利用，即

$$x = \theta + \eta - \alpha r^*$$

其中θ为当前试次决策阈值，η代表随机噪音，服从$N(0, \sigma^2)$的正态分布，r^*代表当前试次的最大奖赏，α代表风险倾向(risk attitude)，当$\alpha > 1$时，越大代表越风险厌恶，即越保守选择利用；当$0 < \alpha < 1$时，越小代表越风险偏好，即越冒险选择探索。

该模型中具有 5 个参数，分别为$k, b, \alpha, \beta, \sigma$。

在原有模型的基础上，基于我们在数据中的观察，我们认为被试在试次间的阈值浮动并非完全随机，还取决于当前试次平均奖赏的情况，根据平均奖赏的大小，被试会动态地调节阈值的高低，从而调节其探索和利用行为的比例。综上，我们提出了模型 2。

4.2 模型 2:Prop-V risk learning Model

我们认为，在经历情境(Experience Context)下的被试，随着试次的增加，被试会逐渐形成对于分布形态的认知。我们使用类似强化学习中时序差分学习(Temporal Difference Learning)的计算形式刻画这一动态过程。假设被试在 *t-1* 时刻存在对于分布的信息

μ_{t-1}, σ_{t-1}。在每个回合结束后，被试会根据本回合新得到的观测值来更新得到μ_t, σ_t，以此来更新分布信息。

$$\mu_t = \mu_{t-1} + \alpha(\overline{x}_t - \mu_{t-1})$$

此处α代表学习率，指被试根据新得到的信息调整自己对于分布均值的认识的程度。$\overline{x}_t$表示该回合中采取探索动作时获得的奖赏的平均值，即

$$\overline{x}_{t-1} = E[\sum_{i=1}^{N} rII(a = 1)]$$

同模型 4，每回合开始前，被试会根据$t-1$时刻的估计值μ_{t-1}来设置阈值

$$\theta_t = \mu_t + k\frac{t_{left}}{T}$$

且阈值设置存在随机噪音η，与$t-1$试次的不确定度有关，即：

$$\eta \propto N(0, \sigma^2)$$

该模型具有 6 个参数，分别为$k, \alpha, \beta, \sigma, \mu_0$和学习率$\alpha$。

经检验，我们发现我们所拟合的模型 2 相较模型 1 在每种条件下都有不同程度的提升，如下图 14 所示。对于某些被试而言，新的模型相较原来的模型有较为显著的提升，说明学习率在试次间的确起到调节作用。而对于某些被试而言，这种提升并不显著，其学习率也接近于 0，即被试采取一种固定阈值进行决策任务。

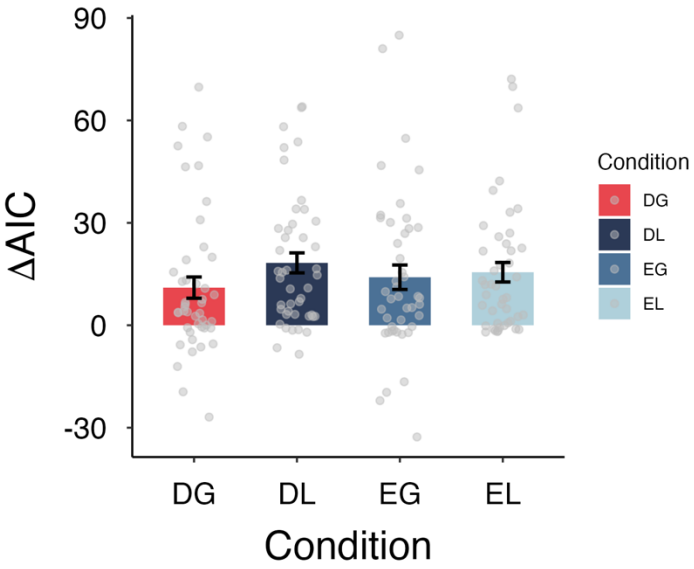

图 16 模型 2 相较于模型 1 在四种条件下拟合优度的提升（通过 AIC 计算）

这种提升在个体间具有较为明显的差异，间接说明被试并不都采用了类似动态阈值调节的策略

第五章 结论与讨论

在本研究中，我们使用了一种极简的探索和利用范式，以剥离掉一些混淆因素例如环境的不确定性，波动性，奖赏动作的结果的未知性等一系列因素的影响，单纯观察人类在探索和利用权衡中的次优性行为的产生，以及情境和效价在其中的调节作用。

首先，我们延续了前人研究中的发现，并发现被试在不同情境下对于环境中变量的敏感性有所差异。例如，在描述情境下的被试会对当前试次的最大奖赏更为敏感，从而根据最大奖赏来调节自己的探索和利用行为。相较而言，经历条件下的被试由于对背后的概率分布未知，因此其行为受到最大奖赏的约束较小。

进一步，我们在原文的基础上进一步挖掘了更多环境变量对于被试行为的影响。我们发现，试次的均值也会对于被试的探索和利用行为产生影响。但是区别于当前试次最大奖赏对于被试探索和利用行为的线性影响，我们发现一部分被试在试次平均奖赏增加时选择会选择继续探索，而一部分则选择利用已有的奖赏。因此，我们认为，被试在其中可能存在一些非理性因素的影响，使得当被试即使获得较好的奖赏时，仍然会继续探索环境。由此，我们引入了 *Gap* 这一变量，并发现在不同水平下 *Gap* 对于探索行为的调节的斜率有所不同，也发现在负 *Gap* 下相比正 *Gap* 下对于探索行为的影响更大，表现出一种不对称性。

除挖掘试次内可能调节被试探索利用行为的影响因素外，我们认为被试在试次间还会具有动态的适应性。具体而言，我们发现被试会根据之前奖赏的增减从而动态地调节当前试次探索动作的多少，且更近的试次对于更远的试次对于当前行为的影响更大，表现为一种近因效应。此外，环境的随机性也会影响当前被试探索行为，即当被试感受到环境的波动较大时，其更倾向于利用当前已有的奖赏。基于以上发现，我们在原有模型的基础上提出了新的模型，引入了试次间的动态调节，并发现该模型相较于原模型的确会有更好的表现。此外，我们还考虑了另外几种备择模型（见**附录 C**），如卡尔曼滤波模型考虑环境的随机性，值分布模型考虑被试如何平衡当前采样奖赏同真实分布之间的差距，非理性 Prop-V risk 学习模型考虑将 *Gap* 作为影响阈值的变量等等。受限于数据和时间限制，这些模型均可以在未来的研究中一一进行测试和评估。

本研究在概念层面上对于心理学研究的贡献在于区分了两种不同的学习和适应性模式，一种可以被称为局部调节(Local adjustment)，指短期内人们受到环境影响从而快速地调节自己的行为模式。另外一种称为全局调节(Global adjustment)，指在一段时间内，人们能够考虑之前与环境交互的情况从而影响当前的决策过程。在我们的研究中，我们的确在试次内和试次间发现了这种效应的存在。在未来的研究中，这一概念可以被作为出发点，设计新的实验范式对其进行验证。事实上，这样二分的概念在心理学研究中层出。例如，环境波动性(Volatility)和随机性(Stochasticity)之间的关系，基于模型的学习(Model-based)和无模型的学习(Model-free)，速度与准确性的权衡(Speed-accuracy tradeoff)，探索和利用的权衡(Exploration-exploitation)，前瞻性和回顾性的规划(Prospective&Retrospective planning)。这些二分的概念间是否存在潜在的联系也是值得被探究的问题。

人类如何提取之前的记忆并且对未来进行规划是目前心理学和神经科学研究中的重要话题。目前有大量的研究都试图解释背后的神经和计算机制。例如，Glascher 等人(2010)发现人脑中可能存在两种不同的学习信号，即状态预测误差(SPE)和奖励预测误差(RPE)，前者主要在顶叶下沟和侧前额叶皮层被表征，是基于模型学习的基础，而 RPE 主要在腹侧纹状体进行表征，是无模型的学习的基础。且随着学习进行，基于模型和无模型的学习对于学习影响的权重会不断发生变化。Daw 等人(2014)年的研究也指出，两种学习方式在学习中各自起到一定作用，在环境变化时，基于模型的学习能够立即更新对于较远状态的预测，而无模型的学习则需要通过经验来调整预测值。

另外一些研究指出，人脑还可能存在对于预测状态的编码，这种预测性的编码可以被看作是基于模型和无模型学习间的一种方式。Mommennejad 等人(2017)年的研究认为人脑采用一种叫做后继表征(Successor Representation)的方式进行学习。Gershman(2018)认为后继表征的规划方式与海马与内嗅皮层的位置细胞和网格细胞的表征存在一定联系。Piray 和 Daw(2021)则提出了一种称为线性强化学习的算法用于解释人脑如何进行规划。他们提出了默认表征的概念，认为我们对于环境进行学习时，如果目标或者环境发生变化，不需要对整体环境结构的表征进行更新，而仅需要更新目标或结构发生变化的区域，这一方式在计算层面可以节省资源。Marcelo 和 Daw(2018)认为人们对于环境的规划和表征是通过强化学习中贝尔曼等式的备份(Back up)来实现的，在考虑接下来所采取的动作时，人们会基于状态本身能够带来的增益(Gain)以及状态的迫近性(Need)来进行选择。

此外，正如本研究中所示，人脑往往有时采用更为简单的启发式进行决策，这种决策的优势可以通过较为节省认知和计算资源的方法得到较为不错的奖赏。另外，人们还可以对环境进行减枝(Prune)，将大的问题拆解成片段(Chunk)进行求解。Sezener 等人(2019)发现动物在学习复杂的运动序列是，可以将运动动作分解来降低计算复杂性，且随着联系的增加，这些组块会逐渐变长。Ramkumar 等人(2016)设计了一种基于速度和准确性的算法用于优化前瞻性规划，该算法可以模拟动物在时间压力下对于规划深度，奖励大小对规划方向的影响，以及在学习和训练过程中如何从目标导向行为转向习惯性行为。Huys 等人发现(2012)人们还可以采取决策树剪枝的策略，即当被试遇到较大损失时会倾向于停止评估，与经典条件反射有关。Van Opheusden 等人(2023)设计了一种棋盘游戏，通过对棋局，反应时和眼动数据的比较，发现人们可以采用一种启发式搜索的计算认知模型。

基于以上对于学习和规划问题讨论，在后续的研究中，研究者可以尝试结合本研究中的新提出的局部和全局调节的概念，并探究其背后更深层次的认知机制和计算原理。此外，启发式决策作为区别于基于模型的学习和无模型学习和决策的第三种决策方法，其背后的神经机制以及与前者之间的联系也上不清楚。是否本任务背后表层的启发式决策蕴含着更深层次的内涵，即随着试次的增加，人们可以在经历条件下学会整体的概率分布，通过表征概率分布来进行决策。后续研究者可以尝试从概念辨析出发，探究其本身背后的神经机制，与已有规划，学习算法间的联系。

参考文献

Bornstein, A. M., Khaw, M. W., Shohamy, D., & Daw, N. D. (2017). Reminders of past choices bias decisions for reward in humans. *Nature Communications*, *8*(1), 15958.

Daw, N. D., & Dayan, P. (2014). The algorithmic anatomy of model-based evaluation. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1655), 20130478.

Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876-879.

Daw, N. D., O'doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, *441*(7095), 876-879.

Garcia, B., Cerrotti, F., & Palminteri, S. (2021). The description–experience gap: a challenge for the neuroeconomics of decision-making under uncertainty. *Philosophical Transactions of the Royal Society B*, *376*(1819), 20190665.

Gershman, S. J. (2018). The successor representation: its computational logic and neural substrates. *Journal of Neuroscience*, *38*(33), 7193-7200.

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.

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in cognitive sciences*, *13*(12), 517-523.

Huys, Q. J., Eshel, N., O'Nions, E., Sheridan, L., Dayan, P., & Roiser, J. P. (2012). Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS computational biology*, *8*(3), e1002410.

Huys, Q. J., Lally, N., Faulkner, P., Eshel, N., Seifritz, E., Gershman, S. J., ... & Roiser, J. P. (2015). Interplay of approximate planning strategies. *Proceedings of the National Academy of Sciences*, *112*(10), 3098-3103.

Kahneman, D., & Tversky, A. (2013). Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I* (pp. 99-127).

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.

Mattar, M. G., & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nature neuroscience*, *21*(11), 1609-1617.

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.

Muller, T. H., Butler, J. L., Veselic, S., Miranda, B., Wallis, J. D., Dayan, P., ... & Kennerley, S. W. (2024). Distributional reinforcement learning in prefrontal cortex. *Nature Neuroscience*, 1-6.

Nassar, M. R., Bruckner, R., & Frank, M. J. (2019). Statistical context dictates the relationship between feedback-related EEG signals and learning. *elife*, *8*, e46975.

Piray, P., & Daw, N. D. (2021). Linear reinforcement learning in planning, grid fields, and cognitive control. *Nature communications*, *12*(1), 4942.

Rouhani, N., Norman, K. A., Niv, Y., & Bornstein, A. M. (2020). Reward prediction errors create event boundaries in memory. *Cognition*, *203*, 104269.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593-1599.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*(5306), 1593-1599.

Sezener, C. E., Dezfouli, A., & Keramati, M. (2019). Optimizing the depth and the direction of prospective planning using information values. *PLoS computational biology*, *15*(3), e1006827.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

van Opheusden, B., Kuperwajs, I., Galbiati, G., Bnaya, Z., Li, Y., & Ma, W. J. (2023). Expertise increases planning depth in human gameplay. *Nature*, *618*(7967), 1000-1005.

Von Neumann, J., & Morgenstern, O. (1947). Theory of games and economic behavior, 2nd rev. Simon, H. A. (1955). A behavioral model of rational choice. *The quarterly journal of economics*, 99-118.

Wulff, D. U., Mergenthaler-Canseco, M., & Hertwig, R. (2018). A meta-analytic review of two modes of learning and the description-experience gap. *Psychological bulletin*, *144*(2), 140. Song, M., Bnaya,

Z., & Ma, W. J. (2019). Sources of suboptimality in a minimalistic explore–exploit task. *Nature human behaviour*, *3*(4), 361-368.

附录 A

附录 A：关于被试在本研究的极简探索和利用范式下的次优性行为的证明(Song et al., 2019)

本研究中的探索和利用任务可以用马尔可夫决策过程(Markov Decision Process)来刻画(Sutton & Barto, 2018)。在原研究描述环境中，被试对收益（损失）分布知情，因此该决策过程看作一种基于模型(Model-based)的学习，可用元组(S, A, P, R)来描述。S 为所有状态 s 的集合，我们认为被试的状态取决于元组(r^*, t_{left})，其中r^*为当前最大的收益，t_{left}为当前试次剩余的天数。A 为每一个状态下可以采取的动作：1（探索）或 0（利用）。转移概率函数$P(s, a, s')$描述了被试处于状态 s 时采取动作 a 转移至状态 s'的概率。若被试在状态 s 下采取利用($a = 0$)时，转移概率被定义为

$$P(s, 0, s') = \begin{cases} 1 & when\ t'_{\text{left}} = t_{\text{left}} - 1\ and\ r^{*\prime} = r*; \\ 0 & otherwise \end{cases}$$

若被试在状态 s 下采取探索($a = 1$)时，转移概率被定义为

$$P(s, 1, s') = \begin{cases} p(r^{*\prime}) & \text{when } t'_{\text{left}} = t_{\text{left}} - 1 \text{ and } r^{*\prime} > r^* \\ Pr(r \le r^{*\prime}) & \text{when } t'_{\text{left}} = t_{\text{left}} - 1 \text{ and } r^{*\prime} = r^* \\ 0 & otherwise \end{cases}$$

$R(s, a, s')$描述在某个状态 s 下采取动作 a 期望所获得的即时奖赏 r，即

$$R(s, a, s') = \begin{cases} r* & a = 0 \\ 3 & a = 1 \end{cases}$$

求解该环境下的最优策略即求解如下贝尔曼方程(Bellman equation)：

$$Q(s, a) = \sum_{s \in S} P(s, a, s')(R(s, a, s') + V(s'))$$

$$V(s') = max\ Q(s, a)$$

$$\pi(a|s) = \text{argmax}\ Q(s, a)$$

经历环境的决策过程可用元组(S, A, R, S)来描述。经历环境中，被试不知道收益（损失）分布的信息，因此可以看作是一种无模型（Model-free）的学习，即被试纯粹依赖与环境交互产生的经历进行学习，我们可用经典的 Q-learning 来刻画被试的学习过程：

$$Q(r*, t_{left};\ a) = Q(r*, t_{left};\ a) + \alpha[(r_{t+1} + \gamma max_{a \epsilon A_{[r*, t_{left}]}} Q(r*, t_{left} - 1; a)) - Q(r*, t_{left}; a)]$$

我们可以同样得到被试在经历环境下的最优策略：

$$\pi(a|s) = \text{argmax}\ Q(s, a)$$

在 Song 等人(2019)的研究中，通过训练智能体在环境中反复学习，其发现在各种试次长度

下，被试在探索和利用之间切换的次数靠近 1，即一旦拿到相对不错的奖赏后则选择一直利用，否则便一直探索下去。在本研究中所设置的 4 种条件下，被试平均切换的次数大于 1，即被试并没有按照理论最优的策略进行决策，如下**图 15A，B** 所示。此外随试次长度的增加，被试在群体水平上获得的平均奖赏增加，如下**图 15C，D** 所示。更细致的理论和数学证明可参考原研究(Song et al., 2019)。

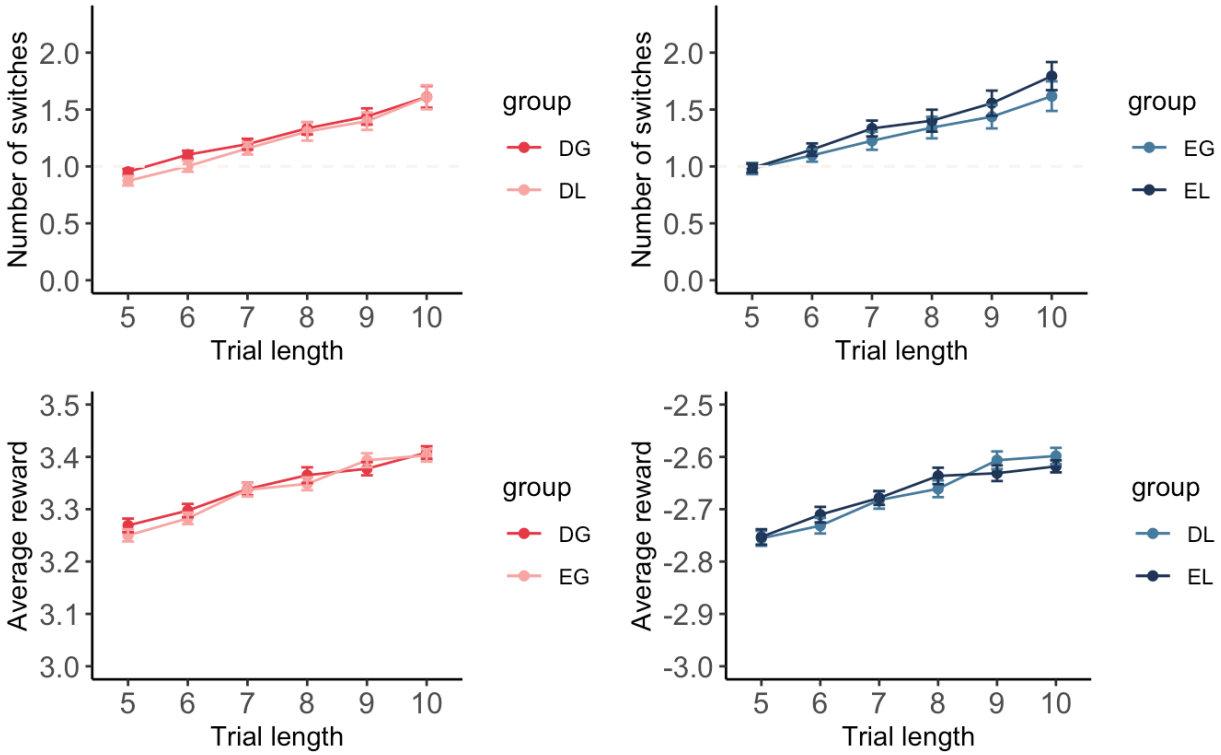

图 15 不同情境和效价条件下随试次长度增加被试在探索和利用之间的切换次数及平均奖赏

A 和 B 分别展示了在描述和经历情境下被试的平均切换次数，与理论最优相偏离。C 和 D 图展示了随试次长度增加，平均奖赏也随之增加。

附录 B

表 2 加入 *Gap* 项后不同情境和效价条件下逻辑斯蒂回归模型拟合的系数值

系数/条件	DG	DL	EG	EL
r^*	-5.860	-5.835	-5.402	-4.761
t_{left}	0.599	0.669	0.824	0.766
T	-0.180	-0.260	-0.297	-0.288
$\bar{r}$	-0.710	-0.701	-0.507	-0.135
gap	0.844	0.710	0.542	1.054
neggap	1.469	1.442	1.361	1.036
gap ×*neggap*	-0.816	-0.566	-0.224	-0.887

表 3 加入 *Gap* 项后不同情境和效价条件下逻辑斯蒂回归模型拟合系数值的显著性 *p* 值

系数/条件	DG	DL	EG	EL
r^*	3.425e-17***	8.384e-14***	3.118e-16***	1.504e-19***
t_{left}	1.745e-15***	1.803e-19***	3.467e-13***	2.732e-16***
T	4.342e-07***	1.705e-09***	6.830e-08***	2.038e-09***
$\bar{r}$	1.54e-02*	5.085e-03**	1.949e-02*	5.981e-01
gap	2.216e-06***	7.898e-06***	4.393e-03**	8.455e-06***
neggap	1.058e-08***	9.011e-09***	1.386e-07***	1.773e-06***
gap ×*neggap*	4.937e-03**	6.985e-03***	3.633e-01	6.344e-04***

注：**p*<0.001, ***p*<0.01, **p*<0.05**

表 4 不同情境和效价条件下三个逻辑斯蒂回归模型的拟合优度

模型/条件	DG	DL	EG	EL
Model1	615.5798	644.4248	651.9644	694.6470
Model2	588.7865	617.5614	636.7952	675.7281
Model3	546.2368	576.3658	593.5462	640.7137

注：拟合优度以 AIC 为标准进行计算

表 5 不同情境和效价条件下上一试次对当前动作调节的回归模型拟合的系数值

系数/条件	DG	DL	EG	EL
$\Delta reward_{t-1}$	-1.389e-03	-9.148e-04	5.959e-03	4.188e-03
$\Delta reward_{t-1}(a=1)$	1.817e-01	1.535e-01	1.472e-01	1.210e-01
Δr^*_{t-1}	3.953e-02	7.004e-02	6.548e-02	5.464e-02
r^*	-2.668e-01	-2.874e-01	-2.931e-01	-2.347e-01
T	-1.782e-02	-2.184e-02	-1.091e-02	-1.774e-02

表 6 不同情境和效价条件下上一试次对当前动作调节的回归模型拟合的系数值的显著性 *p* 值

系数/条件	DG	DL	EG	EL
$\Delta reward_{t-1}$	7.840e-01	8.521e-01	2.409e-01	3.772e-01
$\Delta reward_{t-1}(a=1)$	3.245e-20***	5.174e-16***	2.205e-16***	2.443e-12***
Δr^*_{t-1}	3.959e-03**	4.284e-07***	6.834e-07***	1.163e-05***
r^*	2.447e-19***	5.133e-25***	-7.162e-21***	1.382e-19***
T	6.850e-08***	1.810e-12***	1.977e-04***	1.053e-09***

注：*p<0.001, **p<0.01, *p<0.05***

附录 C

除第四章考虑的两个模型外，我们还考虑了以下几种备择模型，但受限于时间，数据等因素，这些模型在时间内并未能完全拟合成功或表现未能有原有模型表现更好。我们会在之后的研究中继续对其余模型的表现进行一一验证。

模型 **3: Irrational Prop-V risk learning model**

基于对于 *Gap* 的观察，我们认为被试在 *Gap* 为正或负时对于探索行为存在一定不对称性。因此，在模型 3 中，我们参考了卡尼曼的前景理论(Prospect theory)，引入对于 *Gap* 的不对称性，即

$$gap = \begin{cases} -\lambda gap & gap < 0 \\ gap & gap > 0 \end{cases}$$

其中λ代表对于负 *gap* 下的折损。在计算阈值时，我们仍沿用模型 2 中试次间进行阈值更新的公式，即

$$\mu_t = \mu_{t-1} + \alpha(\overline{x}_t - \mu_{t-1})$$

$$\overline{x}_{t-1} = E[\sum_{i=1}^{N} rII(a = 1)]$$

$$\theta_t = \mu_t + k\frac{t_{left}}{T}$$

我们认为在每一次选择中，被试还会受到 *Gap* 的额外影响

$$x = \theta + \eta - \alpha r^* - \rho e^{-r^* gap}$$

这里ρ代表 *Gap* 对于阈值影响的折扣系数，该模型中共有$k, \alpha, \beta, \sigma, \mu_0$和学习率$\alpha$，折扣系数$\rho$和调节正负 *Gap* 参数$\lambda$。

模型 **4: Dynamic Expected Value Prop-V risk Model**（仅适用于 **Description Context)**

在该模型 2 中，我们认为在上述模型 1 中的参数η并不是随机噪音。试次间阈值的调整和被试感知的环境动态性(Volatility)有关。具体而言，当被试的前面试次的整体回报不高时会降低期待，因此降低阈值；整体回报较好时则对应升高期待，增加阈值。

我们记被试在$t-1$试次采取探索动作$(a = 1)$后得到的奖赏均值为$\overline{x}_{t-1}$，即

$$\overline{x}_{t-1} = E[\sum_{i=1}^{N} rII(a = 1)]$$

对于描述(Description)情境下的被试而言，其将该值与真实分布的均值μ进行对比，从而决定下一个试次的阈值，即

$$\eta = \epsilon(\overline{x}_{t-1} - \mu)$$

其中ϵ衡量受到上一试次均值影响的大小

该模型具有 5 个参数，分别为$k, b, \alpha, \beta, \epsilon$

模型 **5: Dynamic Distribution Prop-V risk Model**（仅适用于 **Description Context)**

根据值分布强化学习理论，我们认为被试大脑中存在一群神经元用于负责值分布的表征。具体而言，记环境中的真实分布为$p(x)$，被试通过与环境交互学习的分布为$q(x)$。在决定当前试次的决策阈值时，被试将学习到的分布$q(x)$与真实分布$p(x)$进行对比。我们通过 KL 散度(KL divergence)来衡量真实分布和学习的分布之间的差异：

$$D_{KL}(p||q) = \sum_{i=1}^{n} p(x) log \frac{p(x)}{q(x)}$$

注意 KL 散度恒大于 0，当两个分布完全一致时等于 0，且具有不对称性即$D_{KL}(p||q) \neq D_{KL}(q||p)$。当 KL 散度较大时，代表被试感知的两个分布之间的差异较大，此时意味着环境的不确定性对于被试而言较大。对应更新阈值为：

$$\eta = D_{KL}(p||q)\epsilon(\overline{x}_{t-1} - \mu)$$

该模型具有 5 个参数，分别为$k, b, \alpha, \beta, \epsilon$

模型 **6: Dynamic Kalman Filter Prop-V risk Model**（适用于 **Experience Context)**

我们认为，在经历情境(Experience Context)下的被试，随着试次的增加，被试会逐渐形成对于分布形态的认知。我们使用卡尔曼滤波(Kalman filter)来刻画这一动态过程。

假设被试在$t-1$时刻存在对于分布的先验估计值μ_{t-1}和对于估计值的不确定度σ_{t-1}。在每个回合后，被试通过和环境交互得到一组新的观测值$\overline{x}_t, s_t$，分别代表被试在本回合所有采取探索动作时所获得的均值和标准差，即

$$\overline{x}_t = E[\sum_{i=1}^{N} rII(a = 1)]$$

$$s_t^2 = \frac{1}{N'}\sum_{i=1}^{N'}(\overline{x_t} - r)^2$$

每个回合后，被试会结合自己在$t-1$时刻的估计值和t时刻的新观测值，产生新的估计值μ_t，并校正新估计值的不确定度σ_t

$$\mu_t = \mu_{t-1} + \frac{\sigma_{t-1}^2(\overline{x}_t - \mu_{t-1})}{s_t^2 + \sigma_{t-1}^2}$$

$$\sigma_t^2 = \sigma_{t-1}^2 - \frac{\sigma_{t-1}^4}{s_t^2 + \sigma_{t-1}^2}$$

其中可使：

$$\kappa = \frac{\sigma_{t-1}^2}{s_t^2 + \sigma_{t-1}^2}$$

因此，关于估计值μ_t和不确定度σ_t的更新可写为

$$\mu_t = \mu_{t-1} + \kappa(\overline{x} - \mu_{t-1})$$

$$\sigma_t^2 = (1-\kappa)\sigma_{t-1}^2$$

在每回合开始前，被试会根据$t-1$时刻的估计值μ_{t-1}来设置阈值

$$\theta_t = \mu_t + k\frac{t_{left}}{T}$$

且阈值设置存在随机噪音η，与$t-1$试次的不确定度σ_{t-1}^2有关，即：

$$\eta \propto \sigma_{random}^2 + \lambda\sigma_{t-1}^2$$

其中，λ表示被试对于不确定度的容忍程度。该模型具有 4 个参数，分别为$k, \alpha, \beta, \lambda$

致谢

岁月不居，时节如流。转眼间又到了毕业季。

回想起 4 年前此时正因为疫情在家中备战高考的我，如今已然成为了一名即将踏入科研和学术道路的大学生，这四年实在成长了太多太多。

首先，我最想感谢的便是我的导师李健老师。我曾一度对心理学研究祛魅，甚至对未来感到灰暗和绝望。在我迷茫之际，李老师带我走进了学术的广阔天地，让我颠覆性地改变了对于心理学和神经科学背后意义的认识，并真正感受到了学术思辨和研究的乐趣。李老师对我学术抑或是生活的指导，让我重燃起信心，以一颗平常心面对未来更多更难的挑战。我也想感谢 LiLab 的每一位成员，包括但不仅限于馨茹，哲一，荫梅，章红，杨叟师姐，还有卓凡，许扬，洵伟，嘉澍师兄等。感谢每一位师兄师姐对我的包容，理解和照顾。此外，我也想感谢曾经帮助过我的每一位学术的启蒙老师，包括但不仅限于罗欢老师，张航老师，以及实验室中的各位师兄师姐们，包括但不仅限于穗子，范莹，嘉琪师姐和明浩，东宁，牧之师兄。她们对于学术的严谨和精益求精也督促着我时刻地进步。

第二，我想感谢在大学生涯四年中对我来说不可或缺的一群人。感谢 SEAbling 人声乐团的每一位成员，包括魏来，驭飞，开诚，子萱，愫，雪妍，麒，冠芸，天宇，涵淇，天翼，慧敏，奉时，仕豪，Bella 等。感谢 PKUSO 的每一个人，包括但不仅限于毅为，毅丁，丰驿，荷玥，祎勍，玥彤，越越，伯恒等人。感谢 New Grammar 乐队，包括祥洲，靖怡，林峰，彬洋，悦，虞何，蕊，天翼等人，以及 Spinning Terrasse 乐队的每一个人，包括愫，Felix，成然，喆楷，思危，有为等人。音乐早已成为我生活中不可或缺的一部分，而他们正是我这四年大学生活一直坚持音乐的支柱和源泉。同时这也是我大学四年最志同道合，具有深厚友谊和情谊的一群朋友。回想起每一次专场演出，每一次排练，每一次深夜聊天，这些都将成为我大学四年中最难忘的回忆。感谢我的室友，向渝，基汉，嘉冬，凯然。这大学四年的相处时光，无论是生病时的互相照顾，还是每一次出游团建，都让我记忆犹新。感谢心院 2020 级本科班的朋友们，包括但不仅限于书天，林峰，小钺，玥彤等，他们既是我生活中的好朋友，同样在学习道路上给予我启发和帮助。

第三，我想感谢我的家人，感谢对我一直以来给予我最无条件的爱。毫不夸张地说，大学四年是我有生以来过得最痛苦和最难熬的一段时光，但同时也是我增长见识，重新审视和思考自己未来规划的最关键的一段时光。但不论我的选择是什么，家人总是给予我最可靠的支持，使我没有牵挂和顾虑地去实现自己的理想和抱负。

最后，我想感谢我自己。人脑是世界上最精细和最复杂的系统，其中有太多未知的领域和话题等待着研究者的挖掘和探索。正如文章中所描述的探索和利用一样，人的一生中要做出很多次选择，无论是尝试新鲜的事物或者选择目前自己所拥有的，每一个选择背后都有其背后的价值和意义所在。因此，不必为了一时或一次的选择失利而垂头丧气而欲欲不振，每个人都将通向自己所向往的人生和远方。感谢自己做出的每一个选择。无论未来如何，尽管大步向前。

北京大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

论文作者签名：李佳霖

日期：2024年5月27日

学位论文使用授权说明

本人完全了解北京大学关于收集、保存、使用学位论文的规定，即：

- 按照学校要求提交学位论文的印刷本和电子版本；
- 学校有权保存学位论文的印刷本和电子版，并提供目录检索与阅览服务，在校园网上提供服务；
- 学校可以采用影印、缩印、数字化或其它复制手段保存论文；

论文作者签名：李佳霖 导师签名：李通

日期：2024年5月27日