arXiv:2512.14713v1 [cs.LG] 8 Dec 2025

用于捕捉适应性、反馈驱动出行行为的贝叶斯潜在类别强化学习框架

Georges Sfeir[1], *Stephane Hess*[1], *Thomas O. Hancock*[1], *Filipe Rodrigues*[2], *Jamal Amani Rad*[1], *Michiel Bliemer*[3], *Matthew Beck*[3], *Fayyaz Khan*[4]

[1]利兹大学选择建模中心与交通研究所[2]丹麦技术大学管理学院，丹麦技术大学[3]悉尼大学交通与物流研究所[4]沙特阿拉伯利雅得亚马玛大学商学院

2025年12月18日

摘要

许多出行决策都涉及一定程度的经验形成，即个人会随着时间推移逐步了解自己的偏好。与此同时，个体旅行者之间存在显著差异，不仅潜在偏好各不相同，偏好的演变方式也存在差异。本文提出一种潜在类别强化学习（LCRL）模型，使分析人员能够捕捉这两种现象。我们将该模型应用于驾驶模拟器数据集，并通过变分贝叶斯估计参数。我们识别出三个类别，其成员在偏好适应方式上有显著差异：第一类表现出依情境而变的偏好，并具有情境特定的利用倾向；第二类无论情境如何，都采取持续的利用策略；第三类则采用探索策略，同时具有情境特定的偏好。

关键词

计量经济学；离散选择模型；数学心理学；强化学习；贝叶斯估计

1 引言

理解个人如何在动态且不确定的情境中做出出行决策，是出行行为研究面临的一项根本挑战。传统方法，如离散选择模型（DCM），主要用于将偏好视为固定构念加以捕捉（Cirillo & Xu，2011年），尽管出行者往往会随着新经历的积累而修正自己的预期。在许多现实情境中，个人可能会经历一段适应性学习时期，在此期间，他们会收到反馈（例如意外延误、新出现的备选项），并调整对所面临备选项的预期和偏好。有若干因素可能触发这种基于反馈的适应过程，包括对偏好的出行方式或路线造成干扰或施加临时限制，从而促使出行者探索其他备选项。此外，新的交通备选项——例如新开通的公交线路、共享单车方案或新兴出行服务——也可能促使个人重新考虑惯常的出行模式，并在了解这些备选项的实际表现后更新预期。生活事件和重大的环境变化，包括迁居、开始新工作或搬到另一座城市或国家，也往往需要出行者经历一段探索与适应期，以熟悉可用的交通网络，并找到最符合自身需求的备选项。

由这些多种因素驱动的探索性行为，起初可能会导致数据中观察到的出行选择不稳定且存在差异。然而，随着出行者对所选选项积累经验并增强信心，他们的决策可能逐渐稳定下来，形成由利用驱动的偏好稳定性，并可能产生行为惯性（Rad etal.，2025）。尽管现有研究往往强调对稳定行为模式的识别和量化（Hess & Train，2011年），例如惯性和习惯性行为（Gao 等人，2020；Cantillo 等人，2007年；Cherchi & Manca，2011 年），但通常忽略了这些稳定偏好最初形成的潜在机制。因此，明确建模并全面理解出行者如何通过实际反馈动态学习偏好，仍未得到充分研究（Rad et al.，2025；Henriquez-Jara& Guevara，2025；Henriquez-Jara 等人，2025）。造成这一研究空白的因素有多个，包括缺乏专为捕捉反馈驱动学习而设计的适当纵向数据集；难以通过传统出行调查捕捉动态学习行为，因为这类调查通常只记录选择，却不记录塑造选择的实际反馈或结果；以及传统出行建模方法的局限性。这与数学和行为心理学领域形成鲜明对比，在这些领域中，研究者已开发出多种模型，专门用于研究基于反馈的学习以及偏好随时间的演变（Niv，2009）。

心理学和神经科学中一个成熟的理论是强化学习（RL），它描述了个人如何通过基于经验的反馈来调整预期和行为（Niv，2009；Sutton &Barto，2018）。在这一框架下，数学心理学提供了形式化的强化学习模型，将这一学习过程操作化，以刻画个人的偏好和行为如何根据观察到或接收到的正面或负面反馈随时间演变。在决策情境中，强化学习假设个人会根据经验形成并更新对每个选项／备选项价值的估计。作出选择后，个人会调整所选备选项的预期价值（例如预期出行时间或出行成本），以减少基于实际体验结果产生的预测误差。这种调整或学习过程

由奖励预测误差（RPE）驱动（Collins &Shenhav，2022；Samsonet al.，2010），奖励预测误差定义为个体预期获得的奖励与实际获得的奖励之间的差异。通过反复选择并由此获得反复反馈，这一由奖励预测误差驱动的学习过程使个人能够减少预测误差，强化带来好于预期结果的行为／选择，并抑制导致结果差于预期的行为／选择（Liu etal.，2015）。在强化学习中，利用是指选择当前已学习到的（预期）价值最高的选项；而探索则指有时选择当前价值较低或不确定性较高的选项，例如为了进一步了解其结果并可能改善未来的决策。这类模型已广泛应用于多种行为情境，例如赌博和风险承担（Steingroever et al.，2014）、社会学习（Diaconescu et al.，2017；V´elez & Gweon，2022）、成瘾（Lei et al.，2022）等。

在出行行为研究中，强化学习（RL）能够提供一个框架，用于建模出行者如何根据以往的出行时间、意外延误或新信息，动态更新其偏好。Rad et al.Rad etal.(2025)已证明，强化学习能够准确捕捉两个出行行为数据集中的行为，并揭示出行者偏好背后的学习过程。他们的模型设定通过层级贝叶斯结构纳入异质性，其中个体层面的认知参数取自单成分正态分布或对数正态分布。虽然这种方法允许行为者围绕共同的行为过程呈现连续变化，却无法识别总体中可能存在的不同学习类型（例如，反复尝试不熟悉选项的持续探索者，与迅速收敛到更习惯性行为的快速利用者）。单峰层级先验若不引入额外的混合成分，就无法表示此类多峰性。因此，对这些迥异行为取平均可能导致误导性推断，并使政策干预缺乏针对性（例如，有利于探索者却妨碍利用者的信息策略）。为此，我们提出一种混合式潜在类别强化学习（LCRL）框架，将潜在类别选择模型（LCCM）（例如，见Hess，2024）与类别特定的强化学习选择动态相结合，从而明确建模离散的学习类型。该LCRL模型能够：1）通过将个体划分到具有不同学习动态的潜在类别中，离散地表示未观测异质性（例如，风险厌恶与风险偏好）；2）估计类别特定的动态学习，因为每个潜在类别内部的选择模型都是基于强化学习的决策过程；3）通过类别特定的强化学习模型，明确建模探索-利用权衡行为（即，个体可能通过探索新的备选项获取更多信息，或依据累积的经验利用熟悉的备选项）。此外，该模型采用变分推断在完整的贝叶斯框架下进行估计（Jordan et al.，1999；Wainwright &Jordan，2008），从而为所有模型参数提供完整的后验分布，确保恰当地处理不确定性，并在动态学习与类别划分所带来的高维情况下实现可扩展的估计（Rodrigues，2022）。此外，本研究利用一个独特的数据集，该数据集结合了驾驶模拟器中的重复路线选择与陈述偏好调查的回答。这种设计使出行者能够接收并回应基于实际体验和假设情境的路线选择反馈，从而得以追踪他们的预期如何在真实感和认知投入程度各不相同的情境中演变。本文将所提出的LCRL模型应用于该数据，将数学心理学中的强化学习原理与计量经济学中的潜在类别选择建模相结合，以共同捕捉基于经验的学习与离散形式的

行为异质性。通过这种方式，该模型能够更深入地刻画不同类型的出行者如何学习、探索，并最终稳定其路线选择。

本文其余部分安排如下。首先，我们将在第2节介绍强化学习背后的理论，并阐述所提出的LCRL模型的建模与估计方法论。接着，我们将在第3 节讨论驾驶模拟器路线选择案例研究，并在第4节呈现结果。最后，我们将在第5节作出总结。

2 方法论

本节概述数学心理学中的强化学习（第2.1节）、一个说明性示例（第2.2节），以及所提出的潜在类别强化学习（LCRL）模型的构式（第2.3节）。LCRL模型以一般形式呈现，可应用于出行行为领域，也可能适用于其他领域（例如医疗保健、营销与消费者行为、经济学等）。

2.1 强化学习

数学心理学中的强化学习描述个人如何随着时间推移从反馈中学习。其基本假设是，个人面对选择情境时，会对可供选择或纳入考虑的备选项形成先验预期（Sutton & Barto，2018）。选择某个备选项后，他们会收到某种形式的反馈（结果），根据先前预期与收到的反馈（结果）之间的差异，更新对所选备选项的预期，然后在下一次选择情境中使用更新后的预期。这种基于经验对预期进行更新的过程称为强化学习（RL）。强化学习模型有多种，各自从不同角度阐释个人如何从经验中学习。本文采用Rescorla-Wagner模型（Rescorla &Wagner，1972），该框架解释备选项的价值如何随时间推移而变化，变化依据是预期结果与实际结果之间的差异，这也称为预测误差（Schultz etal.，1997；O'Doherty et al.，2003）。Rescorla-Wagner模型由两个子组件构成：预测误差和选择规则。

2.1.1 预测误差

预测误差 δ 定义为预期结果 Q 与实际结果 r 之差，它会推动未来预期 Q 的更新。因此，个体 $n \in \{1,\dots,N\}$ 对试次 $t \in \{1,\dots,T_n\}$ 中备选项 $i \in \{1,\dots,I\}$ 的期望会随试次逐步变化，如下所示：

$$Q_{nit} = Q_{ni(t-1)} + \alpha\delta_{ni(t-1)} = Q_{ni(t-1)} + \alpha(r_{ni(t-1)} - Q_{ni(t-1)}), \tag{1}$$

其中，$Q_{ni(t-1)}$ 是先验预期，α 是学习率，$\delta_{ni(t-1)}$ 是个体 n在时间 $t-1$对备选项 i产生的预测误差，$r_{ni(t-1)}$ 是备选项 i 在时间 $t-1$的实际结果（反馈）。请注意，与标准的Rescorla–Wagner 仅更新已选择项的表述一致，每个试次仅更新所选备选项的预期。其他强化学习模型允许对未选择的备选项进行虚拟更新，或将未选择项数值的衰减纳入模型，但本构式不考虑这些扩展。此外，Q_{ni0} 表示初始期望

由个体赋予 n 备选项 i在 $t = 0$ 观察到任何结果之前的值。这些初始期望可解释为关于各备选项价值的先验信念，随后会通过经验驱动的反馈进行更新（公式1）。

学习率 α 介于 0 和 1 之间，数值越高，意味着个体在更新其信念或 Q值时越重视近期反馈。这意味着个体会迅速将新信息纳入估计，并大幅降低较早经历的权重。相应地，学习率较低意味着近期结果所占权重较小，预期更新较慢，并保留较早的经历。

2.1.2 选择规则

随后，将预测误差嵌入随机效用最大化（RUM）式框架中，转换为选择概率。具体而言，假设个体 n 在时间 t 选择备选项 i 时的效用取决于个体相应的预期 Q_{nit}。因此，效用表示为：

$$U_{nit} = \gamma_i \pm \beta Q_{nit} + \varepsilon_{nit}, \tag{2}$$

其中， γ_i 是备选项 i的备选项特定常数（或偏差参数）， β 是一个非负的敏感度参数，用于控制个体的探索-利用权衡行为。较高的 β 值会促使个体采取利用行为，而较低的值则会促使个体更多地进行探索。符号（$\pm$）取决于预期结果Q_{nit} 代表奖励（正号）还是成本（负号）。

在这一阶段，假设所有个体具有相同的学习率和敏感度参数。此外，假设所有个体对每个备选项都具有共同的初始期望，记为 $Q_{i,0}$ ，使得$Q_{ni0} = Q_{i,0}$ 对所有个体均成立n。这些简化假设使我们能够在这一阶段聚焦于模型所刻画的一般学习动态，而不引入个体层面异质性。之后通过 LCRL 扩展中的类别特定参数引入异质性。

假设随机误差项 ε_{nit} 在个体和备选项之间独立同分布（iid），且服从 I型极值分布，则个体n 在试次 t 选择备选项 i 的概率遵循传统的多项Logit（MNL）构式，定义如下：

$$P_{nit} \;=\; P\big(y_{nt} = i \;\big|\; \boldsymbol{Q_0}, \boldsymbol{\gamma}, \beta, \alpha\big) \;=\; \frac{\exp\big(\gamma_i \;\pm\; \beta\, Q_{nit}\big)}{\displaystyle\sum_{i'=1}^{I} \exp\Big(\gamma_{i'} \;\pm\; \beta\, Q_{ni't}\Big)}, \tag{3}$$

其中 $\boldsymbol{Q_0} = \{Q_{1,0}, \ldots, Q_{I,0}\}$、 $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_I\}$，其中一个 γ_i 设为零以实现识别，而 $y_{nt} \in \{1, \ldots, I\}$ 则反映个体 n 在情境 t 中的观测选择。随后，个体 n 在所有时间段（情景） T_n 内的全部选择的观测概率 $\boldsymbol{y}_n$ 表示如下：

$$P\big(\boldsymbol{y}_n \mid \boldsymbol{Q_0}, \boldsymbol{\gamma}, \beta, \alpha\big) = \prod_{t=1}^{T_n} P_{ny_{nt}t} \tag{4}$$

2.2 强化学习示例

为更好地说明 Rescorla-Wagner 强化学习框架的运行机制及其与选择建模的联系，考虑一个个体 n，他每天都必须在两条陌生路线 A 和 B 之间做出选择。在第一天/试次（$t = 0$），个体 n 对两个备选项持有相同的先验预期，为每条路线赋予相同的预期出行时间（$Q_{nA0} = Q_{nB0} = 25$）。然而，假设个体对路线 B（$\gamma_B = 1$）的基线偏好高于路线 A（$\gamma_A = 0$）。预期出行时间的敏感度固定为 $\beta = 1$，以在所有试次中平衡探索与利用。由于此处的预期代表出行时间（一种成本），它也以负向形式进入效用函数，反映出对较短出行时间的偏好。在这些假设下，个体 n 在试次 0 选择备选项 A 和 B 时的效用分别为：

- $U_{nA0} = \gamma_A + \beta Q_{nA0} = 0 - 1 \times 25 = -25,$
- $U_{nB0} = \gamma_B + \beta Q_{nB0} = 1 - 1 \times 25 = -24.$

根据公式3，在试次0，个体 n 选择备选项A和B的概率P_{nA0} 和 P_{nB0}分别为0.27和0.73。假设个体 n 选择了备选项B，但由于意外延误，经历的出行时间为30分钟（r_{nB0}）。这导致 $\delta_{nB0} = r_{nB0} - Q_{nB0} = 5$ 分钟的预测误差。这段经历对备选项B未来预期的影响程度取决于学习率（公式1）。表1 展示了在两种不同学习率情境下，试次 $t = 1$的预期和选择概率将如何变化：高学习率（$\alpha = 0.9$）和低学习率（$\alpha = 0.1$）。

表1：$t = 1$时的预期和概率

$t = 1$	高学习率（$\alpha = 0.9$）	低学习率（$\alpha = 0.1$）
Q_{nA1}	25（未更新）	25（未更新）
Q_{nB1}	$25 + 0.9 \times 5 = 29.5$	$25 + 0.1 \times 5 = 25.5$
U_{nA1}	-25	-25
U_{nB1}	-28.5	-24.5
P_{nA1}	0.97	0.38
P_{nB1}	0.03	0.62

如果个体 n 的学习率较高（$\alpha = 0.9$），就会迅速修正对先前所选备选项的期望，导致路线 B 的效用急剧下降（U_{nB1} =-28.5），其选择概率也随之骤降（$P_{nB1} = 0.03$）。在这种情况下，个体 n 会强烈倾向于在 $t = 1$转向路线 A。相反，学习率较低（$\alpha = 0.1$）时，对路线 B 的期望会保持在接近原始值的水平（$Q_{nB1} = 25.5$），其效用（U_{nB1} =-24.5）和选择概率（$P_{nB1} = 0.62$）只会小幅变化，因此路线 B 仍是最具吸引力的备选项。在这种情况下，个体 n 仍会倾向于选择最初偏好的路线 B，同时逐渐积累新信息。这个例子说明了强化学习模型如何捕捉个人对反馈的反应。较高的学习率会促使个体根据新的经历迅速适应，从而增加探索性行为，并导致对近期经历反应过度。相反，较低的学习率会使个体对近期反馈的适应较为缓慢但逐步进行，从而强化习惯性或先前偏好的选择，并减少对近期负面反馈的过度反应。

2.3 潜在类别强化学习（LCRL）模型 l

所提出的模型将计量经济学中的潜在类别（LC）模型与数学心理学中的类别特定强化学习（RL）相结合，旨在通过潜在类别捕捉不同的行为模式，并利用各类别内的强化学习处理决策中的动态反馈。潜在类别模型广泛用于计量经济学，以表示离散形式的异质性；强化学习则用于数学心理学，以描述基于经验的学习，但将二者整合到一个统一框架中的情况仍然少见。据我们所知，这是首次采用完整的贝叶斯变分方法，对离散行为异质性与动态强化学习进行联合建模。因此，这种混合型潜在类别强化学习（LCRL）模型能够识别定性上不同的学习类型，例如探索型与利用型决策者，同时考虑每个潜在类别内偏好的时间演变。

2.3.1 潜在类别成员资格

潜在类别成员概率反映了个体 n 属于类别 $k \in \{1, \ldots, K\}$ 的可能性，该概率取决于个人的社会人口特征和社会经济特征，以及情境属性。这些概率由以下定义的成员函数得出：

$$I_{nk} = \boldsymbol{X}_n^\top \boldsymbol{\eta}_k + \upsilon_{nk}, \tag{5}$$

其中， $\boldsymbol{X}_n$ 是个体 n 的特征及情境属性（包括常数项）构成的向量， $\boldsymbol{\eta}_k$ 是相应的未知参数向量， υ_{nk} 是随机扰动项。为实现识别，将某一类别的参数设为零，例如 $\eta_K = 0$。

假设扰动项在个人和类别之间独立同分布，且服从 I 型极值分布，则个体 n 属于类别 k 的概率定义如下：

$$P\big(s_n = k|\boldsymbol{X}_n, \boldsymbol{\eta}\big) = \frac{\exp\big(\boldsymbol{X}_n^\top \boldsymbol{\eta}_k\big)}{\sum_{j=1}^{K} \exp\big(\boldsymbol{X}_n^\top \boldsymbol{\eta}_j\big)}, \tag{6}$$

其中， $s_n \in \{1, \ldots, K\}$ 和 $\boldsymbol{\eta} =$ （$\boldsymbol{\eta}_1$、……、 $\boldsymbol{\eta}_K$）。

2.3.2 类别特定的强化学习选择模型

所提出模型的类别特定选择成分遵循第 2.1 节介绍的强化学习结构，假设 Rescorla-Wagner模型的两个子成分——预测误差和选择规则——具有类别特异性。为完整起见，我们在下文重述类别特定版本的强化学习选择模型。尽管该模型与第 2.1 节中的标准强化学习构式非常相似，但通过显式标注类别索引来呈现方程，有助于引入后续生成过程所用的符号，并确保不太熟悉强化学习模型的读者也能清楚理解。

预测误差

在每个类别中，个体对备选项 i 在试次 t时的期望，在给定其属于类别 k的条件下，将随试次推移而变化，具体如下：

$$Q_{nitk} = Q_{ni(t-1)k} + \alpha_k \delta_{ni(t-1)k} = Q_{ni(t-1)k} + \alpha_k \Big(r_{ni(t-1)} - Q_{ni(t-1)k} \Big), \tag{7}$$

其中， α_k 是属于类别 k的个人的学习率， $\delta_{ni(t-1)k}$ 是个人 n 在时间 $t-1$对备选项 i 所产生的预测误差，而$r_{ni(t-1)}$ 是个人 n在时间 $t-1$经历的备选项 i的实际结果（反馈）。请注意，初始期望 Q_{i0k} 被视为自由参数，并根据观测数据进行估计，而不是事先设定。

选择规则

在每个潜在类别内，预测误差学习过程通过第2.1.2节介绍的RUM风格框架转换为选择。预期值 Q_{nitk}被视为类别特定的效用的系统性组成部分，并与数学心理学中的两个类别特定认知参数相结合，即偏差 γ_{ik} 以及探索—利用 β_k 参数。在给定类别 k的条件下，个人 n在试次 i 选择备选项 t 的概率为：

$$P_{nitk} = P\big(y_{nt} = i | \boldsymbol{Q}_{0k}, \boldsymbol{\gamma}_k, \beta_k, \alpha_k, s_n = k\big) = \frac{\exp\big(\gamma_{ik} \pm \beta_k Q_{nitk}\big)}{\sum_{i'=1}^{I} \exp\Big(\gamma_{i'k} \pm \beta_k Q_{ni'tk}\Big)}, \tag{8}$$

其中， $\boldsymbol{Q}_{0k} = \{Q_{10k}, \ldots, Q_{I0k}\}$和 $\boldsymbol{\gamma}_k = \{\gamma_{1k}, \ldots, \gamma_{Ik}\}$表示每个类别对应的一个 γ_{ik}；为实现识别，对于一个备选项 i，每个类别的 k均归一化为零； $y_{nt} \in \{1, \ldots, I\}$ 则捕捉了观察到的选择。

给定类别 k，个体 n 在所有时间段（情景） T_n 内观察到全部选择的概率表示如下：

$$P\big(\boldsymbol{y}_n | \boldsymbol{Q}_{0k}, \boldsymbol{\gamma}_k, \beta_k, \alpha_k, s_n = k\big) = \prod_{t=1}^{T_n} P_{nitk}. \tag{9}$$

2.3.3 BayesianLCRL：联合模型与似然

令 $\boldsymbol{\Theta} = \{\boldsymbol{s}, \boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{Q_0}\}$ 表示所提出的 LCRL 模型中所有需要估计的备选项和/或类别上的潜变量集合。个体 n观测选择的无条件概率（或似然），可通过对条件强化学习选择概率（公式 9）按其属于各类别的概率 k （公式 6）进行混合而得。然后，假设有一个由独立个体构成的样本，所有 N 个体的似然为：

$$P(\boldsymbol{Y} | \boldsymbol{\Theta}) = \prod_{n=1}^{N} \sum_{k=1}^{K} P\big(s_n = k | \boldsymbol{X}_n, \boldsymbol{\eta}\big) \; P\big(\boldsymbol{y}_n | \boldsymbol{Q}_{0k}, \boldsymbol{\gamma}_k, \beta_k, \alpha_k, s_n = k\big), \tag{10}$$

其中， $\boldsymbol{Y} = \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ 是所有个体的观测选择集合。

所有未知参数和潜变量均以完整贝叶斯方式指定。完整贝叶斯LCRL模型的生成过程可概述如下：

1. **对于每个潜在类别** $k \in \{1, \ldots, K-1\}$:

   (a) 抽取类别分配参数 $\boldsymbol{\eta}_k \sim \mathcal{N}\big(\mu_{\boldsymbol{\eta}_k}, \ \Sigma_{\boldsymbol{\eta}_k}\big)$

2. **对于每个潜在类别** $k \in \{1, \ldots, K\}$:

   (a) 抽取类别特定的强化学习探索-利用参数 $\beta_k \sim$ 对数正态分布$\big(\mu_{\beta_k}, \ \sigma_{\beta_k}\big)$，以确保 $\beta_k > 0$。(b) 抽取类别特定的强化学习学习率参数 $z_{\alpha_k} \sim \mathcal{N}\big(\mu_{z_{\alpha_k}}, \ \sigma_{z_{\alpha_k}}\big)$，然后应用 Sigmoid 变换，以确保学习率处于 $[0, 1]$: $\alpha_k = S\big(z_{\alpha_k}\big) = \frac{1}{1+e^{-z_{\alpha_k}}}$(c) **对于每个备选项** $i \in \{1, \ldots, I\}$: i. 抽取类别特定的强化学习偏差参数 $\gamma_{ik} \sim \mathcal{N}\big(\mu_{\gamma_{ik}}, \ \sigma_{\gamma_{ik}}\big)$（出于识别目的，每个类别中的一个备选项设为零。）ii. 抽取初始期望 $z_{Q_{i0k}} \sim \mathcal{N}\big(\mu_{z_{Q_{i0k}}}, \ \sigma_{z_{Q_{i0k}}}\big)$，然后应用以下 Sigmoid 变换，以确保初始$Q$ 值为正且处于 $[a, b]$: $Q_{i0k} = a + (b-a) \times S\big(z_{Q_{i0k}}\big)$，其中 a和 b为固定值，由实验设计确定，不进行估计。

3. **对于每个个体** $n \in \{1, \ldots, N\}$:

   (a) 抽取潜在类别分配 $s_n \sim$ 多项Logit$\big(\boldsymbol{X_n}, \ \{\boldsymbol{\eta}_k\}_{k=1}^K\big)$

   (b) **对于每个潜在类别** $k \in \{1, \ldots, K\}$:

      i. **对于每个选择时点** $t \in \{1, \ldots, T_n\}$:

         A. Draw observed choice $y_{nt} \sim \text{MNL}\big(\beta_k, \{\gamma_{ik}, Q_{nitk}\}_{i=1}^I\big)$,

         B. Compute the prediction error of the chosen alternative $y_{nt} = j$ (according to step A): $\delta_{njtk} = r_{njt} - Q_{njtk}$

         C. Update the expectation of the chosen alternative $y_{nt} = j$ (according to step A): $Q_{nj(t+1)k} = Q_{njtk} + \alpha_k \, \delta_{njtk}$

给定贝叶斯生成过程，LCRL模型的联合分布定义如下:

$$
\begin{aligned}
P(\boldsymbol{Y}, \boldsymbol{\Theta}) = & \prod_{k=1}^{K-1} P(\boldsymbol{\eta}_k) \left[\prod_{k=1}^{K} P(\beta_k) P(\alpha_k) \left(\prod_{i=1}^{I-1} P(\gamma_{ik}) \right) \prod_{i=1}^{I} P(Q_{i0k}) \right] \\
& \times \prod_{n=1}^{N} \prod_{k=1}^{K} P(s_n = k | \boldsymbol{X}_n, \boldsymbol{\eta}) P(\boldsymbol{y}_n | \boldsymbol{Q}_{0k}, \boldsymbol{\gamma}_k, \beta_k, \alpha_k, s_n = k), \qquad (11)
\end{aligned}
$$

因此，我们的目标是根据观测数据推断 $\boldsymbol{\Theta}$ 。根据贝叶斯定理， $\boldsymbol{\Theta}$的后验分布为:

$$
P(\boldsymbol{\Theta} | \boldsymbol{Y}) = \frac{P(\boldsymbol{Y} | \boldsymbol{\Theta}) P(\boldsymbol{\Theta})}{\int P(\boldsymbol{Y} | \boldsymbol{\Theta}) P(\boldsymbol{\Theta}) \, d\boldsymbol{\Theta}}, \qquad (12)
$$

其中，对 $\boldsymbol{\Theta}$ 的依赖关系是显式的。然而，由于分母（即边际似然）涉及高维积分，该后验分布在计算上难以处理。因此，需要采用近似推断方法。

2.3.4 变分推断

LCRL模型推断的目标是估计驱动个体决策的潜变量和未知参数。这些参数包括潜在类别分配 s_n、将个体观测属性与潜在类别成员资格联系起来的类别成员资格参数 η_k 、类别特定的强化学习参数——即偏差参数 γ_{ik} 和探索—利用参数 β_k——以及通过潜变量 z_{α_k}参数化的学习率参数，还有初始Q值参数 $z_{Q_{i0k}}$。

鉴于真实后验分布 $P(\boldsymbol{\Theta}|\boldsymbol{Y})$在计算上难以处理，我们依据变分推断（VI）或变分贝叶斯（VB）的原理对其进行近似（Jordan etal.，1999；Wainwright& Jordan，2008），方法是选取一族易于处理的分布 $q_\phi(\boldsymbol{\Theta}$，其参数由 ϕ表示。由此，推断问题转化为一个优化问题，目标是确定参数 ϕ ，使变分近似 $q_\phi(\boldsymbol{\Theta})$尽可能接近真实后验分布 $P(\boldsymbol{\Theta}|\boldsymbol{Y})$。在此框架下，我们针对潜在类别分配 $\boldsymbol{s}$ 和潜变量 $\{\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \boldsymbol{Q_0}\}$引入相互独立的变分分布。为使推断易于处理，我们采用平均场变分近似，并假设变分后验在所有潜变量上因子化：

$$q_\phi(\boldsymbol{\Theta}) = \prod_m q_\phi(\Theta_m), \tag{13}$$

这意味着在变分近似中， $\boldsymbol{\Theta}$ 的所有组成部分之间条件独立。类别归属 $q_\phi(s_n)$的变分分布采用类别分布建模，其参数为 π_n，表示个体n 属于潜在类别 k的概率。类别成员资格参数 $q_\phi(\eta_k)$的变分分布采用多元正态分布建模，其均值向量为 μ_{η_k}，协方差为 $\sum_{\eta_k}$。类似地，类别特定的强化学习偏差参数 $q_\phi(\gamma_{ik})$的变分分布采用正态分布建模，并以均值向量$\mu_{\gamma_{ik}}$ 和标准差$\sigma_{\gamma_{ik}}$为参数。类别特定的强化学习探索-利用参数 $q_\phi(\beta_k)$的变分分布采用对数正态分布建模，参数为 μ_{β_k} 和 σ_{β_k}，从而确保 β_k 为正。最后，驱动学习率动态变化的潜变量 z_{α_k} 以及初始 Q-值 $z_{Q_{i0k}}$ 分别采用正态分布 $q_\phi(z_{\alpha_k})$和 $q_\phi(z_{Q_{i0k}})$建模。随后，通过Sigmoid函数对这些潜变量进行变换，以确保得到的学习率和初始Q-值处于有效范围内。

真实后验分布与近似后验分布之间的距离可以用 Kullback–Leibler（KL）散度来度量：

$$\mathrm{KL}\big(q_\phi(Z) \,\|\, P(\boldsymbol{\Theta} \mid \boldsymbol{Y})\big) = \int q_\phi(\boldsymbol{\Theta}) \log\left(\frac{q_\phi(\boldsymbol{\Theta})}{P(\boldsymbol{\Theta}|\boldsymbol{Y})}\right) d\boldsymbol{\Theta} \tag{14}$$

然而，上述 KL 散度难以计算，因此无法直接最小化。相反，根据 VI 理论（Jordan et al.，1999；Wainwright& Jor-dan，2008），可以通过最大化证据下界（ELBO）来等价地优化变分参数，其定义如下：

$$\mathrm{ELBO} = \mathbb{E}_{q_\phi}\left[\log P(Y|\boldsymbol{\Theta})\right] - \mathrm{KL}\big(q_\phi(\boldsymbol{\Theta}) \,\|\, P(\boldsymbol{\Theta})\big) \tag{15}$$

其中 $\mathbb{E}_q[\log P(Y \mid \boldsymbol{\Theta})]$ 是模型的期望对数似然，而 $\mathrm{KL}\big(q_\phi(\boldsymbol{\Theta}) \parallel P(\boldsymbol{\Theta})\big)$是 Kullback–Leibler散度，用于正则化所有未 $\boldsymbol{\Theta}$ 知参数的变分分布，使其接近各自的先验分布。结合模型的联合分布（式11），以及针对 $q_\phi(\boldsymbol{\Theta})$ 所采用的完全独立（均值场）近似，KL 散度项可分解如下：

$$\begin{aligned}\mathrm{KL}(q_\phi(\boldsymbol{\Theta}) \parallel P(\boldsymbol{\Theta})) = \mathrm{KL}(q_\phi(s_n) \parallel P(s_n)) + \mathrm{KL}(q_\phi(\eta_k) \parallel P(\eta_k)) + \mathrm{KL}(q_\phi(\beta_k) \parallel P(\beta_k)) \\ + \mathrm{KL}(q_\phi(\gamma_{ik}) \parallel P(\gamma_{ik})) + \mathrm{KL}(q_\phi(z_{\alpha_k}) \parallel P(z_{\alpha_k})) \\ + \mathrm{KL}(q_\phi(z_{Q_{i0k}}) \parallel P(z_{Q_{i0k}}))\end{aligned} \tag{16}$$

针对变分参数最大化证据下界 ϕ 等价于最小化式 14 中的 KL 散度。我们使用Adam优化器（Kingma& Ba，2017）优化证据下界。Adam 是一种基于随机梯度的方法，可在高维情形下高效估计变分参数。如式 15 所示，这种方法在拟合观测数据与模型复杂度之间取得平衡，并利用先验所提供的正则化。我们采用基于随机梯度的优化方法（Kingma &Ba，2017；Hoffman etal.，2013）最大化证据下界，从而联合推断观测数据背后的潜在类别结构与强化学习动态。

请注意，第2.1 节中的强化学习模型也采用变分推断进行估计，其生成过程见附录A。强化学习模型和所提出的 LCRL 模型均使用 Python 和 PyTorch （ Paszke et al.，2019）实现。此外，附录B还提供了这两种模型的参数恢复分析。此验证步骤评估估计程序从合成数据中恢复已知参数的能力，以确保所提出的变分推断方法能够得出可靠且可识别的估计值。

3 数据集

所提出的 LCRL 模型在悉尼大学收集的一组驾驶模拟器数据集上进行了测试（ Fayyaz，2021年）。103 名参与者完成了 20 项选择任务，需要在一条出行时间固定（始终为 5 分钟）的路线（下称可靠路线）和一条出行时间不确定、存在变化的路线（他们被告知该路线的出行时间可能在 2 至 7 分钟之间，下称不可靠路线）之间做出选择。实验分为两个主要阶段，两阶段间隔一至两周。

在第一部分中，每位参与者都前往悉尼大学的实验室，并在驾驶模拟器（DS）中进行10次选择，亲身体验出行时间。如果他们选择可靠路线，则驾驶时间恰好为5分钟。如果他们选择不可靠路线，则驾驶时间将从预先指定的2分钟或7分钟分布中抽取，其中抽到2分钟的概率为60%，抽到7分钟的概率为40%（因此，平均出行时间为4分钟）。重要的是，该出行时间分布对参与者而言是未知的，参与者只有在完成行程后才能得知不可靠路线的实际持续时间，并可通过反复选择了解出行时间分布。不可靠路线的出行时间抽取结果因参与者而异：有些参与者最初获得较短的出行时间（2分钟），而另一些参与者最初获得较长的出行时间（7分钟）。

在第二部分中，每位参与者完成了一份在线问卷，其中包含10项针对相同的可靠路线和不可靠路线的陈述偏好（SP）选择任务；参与者无需实际驾驶路线，但会立即收到所选路线实际持续时间的书面屏幕反馈。

参与者被随机分配为先进行驾驶模拟器任务或SP任务，以减轻顺序效应。部分参与者在所有情境中都选择了相同路线，因此被从数据集中剔除。最终数据集包含来自83名个人的1,660个选择观测值。在这些参与者中，约49%为女性，50%未满40岁，59%有全职工作，47%先进行驾驶模拟器任务（表2）。

图1 展示了整个样本以及各主要社会人口、经济和实验特征群体中，可靠路线与不可靠路线选择所占的比例。总体而言，参与者选择不可靠路线（61.7%）的频率高于可靠路线（38.3%）。女性在46.34%的选择中选了可靠路线，高于男性；男性仅有30.4%的选择选了该路线。因此，男性对不可靠路线的偏好似乎比女性更强（分别为69.6%和53.7%）。不同年龄组的路线偏好差异不大，不过，50–59岁的参与者选择可靠路线的倾向最高（40.24%）。从收入水平来看，来自较高收入家庭（10万澳元及以上）的参与者选择可靠路线的频率略高于较低收入群体。就教育水平而言，拥有研究生学位或更高学历的参与者选择不可靠路线的频率（53.5%）低于教育水平较低的群体（超过60%）。实验情境对路线选择有显著影响。参与者在驾驶模拟器任务中选择不可靠路线的倾向（66.5%）高于在SP任务中的倾向（57%）。此外，先进行SP任务的参与者比先进行驾驶模拟器任务的参与者更倾向于选择可靠选项（分别为45.4%和31.9%）。最后，就业状况和驾照类型对各群体选择差异的影响均不明显。

图2 通过展示路线选择计数（可靠路线与不可靠路线），并进一步按实验条件的顺序（DS 优先与 SP 优先）和任务类型（DS 与 SP任务）进行细分，提供了更多见解。先进行驾驶模拟器任务的参与者在 DS 阶段（293 次选择不可靠路线，147 次选择可靠路线）和后续的 SP任务中（306 次选择不可靠路线，134 次选择可靠路线）始终偏好不可靠路线，这表明初次接触体验式反馈显著提高了参与者的风险容忍度。相比之下，先进行 SP任务的参与者最初偏好可靠路线（223 次选择可靠路线，167 次选择不可靠路线）。然而，当这些参与者之后使用驾驶模拟器驾驶时，他们的偏好转向不可靠路线（259 次选择不可靠路线，131 次选择可靠路线），凸显了实际反馈在塑造路线选择方面的重要作用。

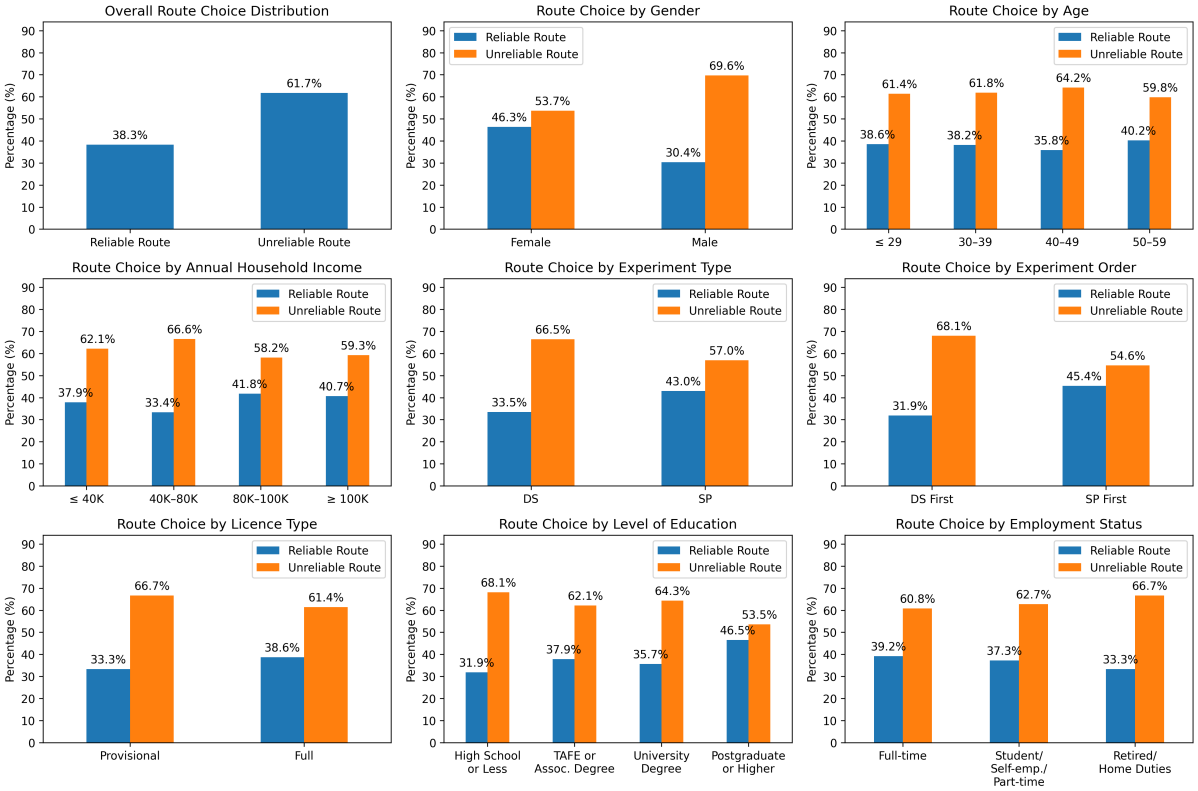


图1：不同人口统计特征和实验特征下的路线选择分布

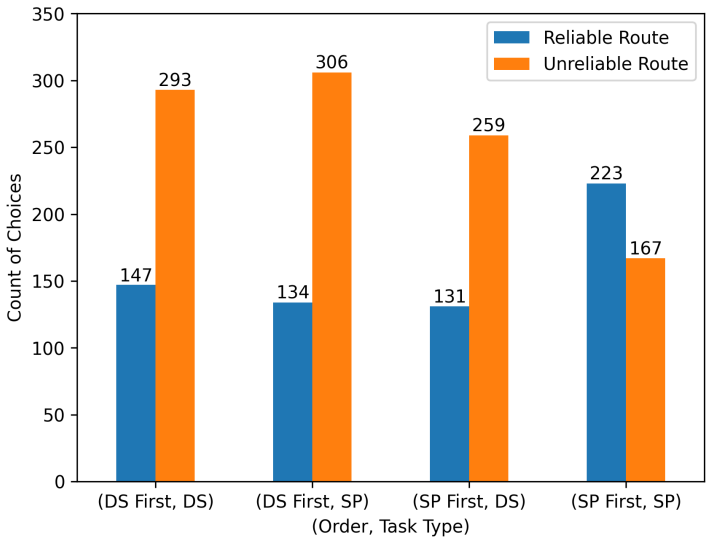


图2：按顺序（DS 优先与 SP 优先）和任务类型（DS 与 SP）划分的选择次数

4 结果

鉴于上一节所呈现数据的特点，第2.2 节介绍的潜在类别强化学习模型据此进行了调整，以反映

表2：不同人口统计特征和实验特征下的参与者人数及路线选择次数

		参与者人数		选择次数	
		计数	%	可靠路线	不可靠路线
总数	—	83	—	635	1025
性别	女性	41	49.40%	380	440
	Male	42	50.60%	255	585
Age	1：29岁或以下	21	25.30%	162	258
	2：30至39岁	22	26.51%	168	272
	3：40至49岁	19	22.89%	136	244
	4：50至59岁	21	25.30%	169	251
收入	1: 40 ,000或以下	7	8.43%	53	87
	2: 40 ,001 至 80 ,000	28	33.73%	187	373
	3: 80 ,001 至 100 ,000	19	22.89%	159	221
	4：超过 100 ,000	29	34.94%	236	344
实验类型	驾驶模拟器（DS）	83	100.00%	278	552
	陈述偏好（SP）	83	100.00%	357	473
实验顺序	DS 优先	39	46.99%	354	426
	SP 优先	44	53.01%	281	599
许可证类型	2：P2	6	7.23%	40	80
	3：完整	77	92.77%	595	945
教育水平	1：高中及以下	8	9.64%	51	109
	2：副学士学位、文凭	28	33.73%	212	348
	3：大学学位	30	36.14%	214	386
	4：研究生及以上	17	20.48%	158	182
就业类型	1：全职	49	59.04%	384	596
	2：学生、兼职、自雇	31	37.35%	231	389
	3：退休或家务	3	3.61%	20	40

参与者的具体决策环境，并有效捕捉多样化的学习动态和初始路线选择预期。研究人员告知参与者，可靠路线的出行时间始终为 5 分钟。因此，所有个人和类别对可靠路线的初始期望$Q_{Rel,\ 0,\ k}$均设定为 5 分钟。相反，研究人员告知参与者，不可靠路线的出行时间将在 2 至 7 分钟之间变化。因此，依据第$Q_{Unr,\ 0,\ k}$节生成过程第 2.(c).ii 步中概述的 Sigmoid 变换，不可靠路线的初始期望采用概率方式建模，其范围边界分别对应于参数2.2 $a = 2$ 和 $b = 7$，从而确保 $Q_{Unr,\ 0,\ k}$ 落在该范围内。

此外，为更好地反映数据集的实证特征，对敏感度参数 β_k 进行了修订。描述性分析（第3节）显示，参与者在驾驶模拟任务和SP任务期间做出了不同的路线选择，因为他们在驾驶模拟任务中往往更具探索性或更偏好风险（图1）。为了反映参与者的探索—利用行为可能因其是在实际体验条件下还是在假设情境下做出选择而存在显著差异，现分别引入驾驶模拟器（$\beta_{DS,\ k}$）和陈述偏好任务（$\beta_{SP,\ k}$）的类别特定敏感度参数。具体而言，敏感度参数定义如下：

$$\beta_k = \beta_{DS,k} D + \beta_{SP,k}(1 - D), \tag{17}$$

其中，D 是一个虚拟变量：如果观测值对应驾驶模拟器，则取值为 1；否则取值为 0。

类似地，偏差参数 γ_{ik} 定义如下，用于捕捉数据集经验特征的影响：

$$\gamma_{ik} = \gamma_{DS,i,k} + \gamma_{SP,i,k}(1 - D), \tag{18}$$

其中，$\gamma_{DS,i,k}$ 表示驾驶模拟器实验中备选项 i 的基线偏差/偏好，而 $\gamma_{SP,i,k}$ 是一个偏移参数，表示其相对于驾驶模拟基线的偏离

在陈述偏好情境中。

根据第3节的描述性分析，我们将类别成员资格模型设定为关键社会经济/人口统计特征和实验特征的函数。具体而言，类别成员资格受性别、年龄（未满40岁编码为1，否则编码为0）、收入（家庭年收入低于80,000澳大利亚元编码为1，否则编码为0）、教育水平（参与者持有研究生学位或更高学位编码为1，否则编码为0）以及实验顺序（参与者先完成驾驶模拟器任务还是陈述偏好任务）的影响，此外还包括一个常数项。

我们估计了最多包含四个类别的LCRL模型，并以标准强化学习模型（如第2.1节所述）作为基准。为确保两种模型之间的比较公平且一致，标准强化学习模型也采用了相同的初始期望、偏差和敏感度参数设定。表3报告了强化学习模型和LCRL模型的拟合优度统计量。基准强化学习模型（K = 1）的对数似然（LL）最低，赤池信息准则（AIC）和贝叶斯信息准则（BIC）值最高。随着类别数量增加，模型拟合有所改善，其中四类别LCRL模型的LL和AIC最佳。然而，三类别模型的BIC最低，因此被选为后续分析的首选设定。

表3：模型比较统计量

K	参数数量	LL	AIC	BIC
1 - RL基线模型	6	-962.91	1,937.82	1,970.31
2	18	-870.16	1,776.32	1,873.78
3	30	-803.51	1,667.02	**1,829.46**
4	42	**-786.23**	**1,656.46**	1,883.87

表4 展示了三类LCRL模型和标准RL模型的结果。对于每个参数，我们报告后验均值，并在括号中给出近似后验z分数，其定义为后验均值与后验标准差之比。如前所述，LCRL模型在对数似然、赤池信息准则和贝叶斯信息准则方面均显著优于RL基准模型。这些改进体现了通过潜在类别划分来考虑学习行为和偏好中的未观测异质性的益处。LCRL模型识别出三个不同的潜在类别，可按以下方式命名和描述。

第*1*类 - 具有依赖情境的偏好和利用倾向的*SP*优先参与者：第1类约占样本的22%。首先接触SP任务的参与者属于该类别的后验概率更高（“驾驶模拟器优先”系数为负且显著）。其余类别归属系数均不显著。该类别成员表现出强烈的依情境而变的偏好。他们对驾驶模拟器中的可靠路线表现出统计上显著的厌恶（$\gamma_{DS,\ Rel,\ k}$ = -0.635且显著），这表明当结果通过实际体验获得时，他们更偏好不可靠路线。然而，较大且显著的正SP偏移参数（$\gamma_{SP,\ Rel,\ k}$ = 3.70）表明，在假设性SP条件下，这一偏好发生了明显逆转，参与者明显更倾向于可靠路线。两个敏感度参数均为正（$\beta_{DS,\ k}$ = 0.935和$\beta_{SP,\ k}$ = 0.781），但在统计上不显著，表明选择并未受到所捕捉的探索—利用机制的强烈支配，该机制体现于

表4：基线和LCRL结果

变量	RL	LCRL		
		第1类	第2类	第3类
	强化学习选择模型	类别特定强化学习选择模型		
$\gamma_{\mathrm{DS},Rel,k}$	$-0.799\,(-14.81)$	$-0.635\,(-4.00)$	$-1.49\,(-15.87)$	0.372 (3.88)
$\gamma_{\mathrm{SP},Rel,k}$	0.299 (2.77)	3。70（8。37）	0.390 (2.95)	$-0.684\,(-5.17)$
$\beta_{\mathrm{DS},k}$	0.419 (5.81)	0.935 (0.40)	0.337 (3.42)	0.247 (4.27)
$\beta_{\mathrm{SP},k}$	1.00 (0.03)	0.781 (0.62)	0.837 (1.11)	0.141 (4.55)
α_k	0.251 (8.32)	0。277（3。55）	0.355 (2.19)	0.437（0.42）
$Q_{\mathrm{Unr},0,k}$	6.69 (8.25)	6。16 (5。31)	6.12 (4.05)	4.81 (0.36)
		类别成员资格		
η_{ASC}		$-0.0403\,(-0.08)$	0.0291 (0.06)	
$\eta_{\mathrm{DS第一}}$		$-1.02\,(-2.10)$	0。943 (2。53)	
η_{Female}		$-0.121\,(-0.27)$	$-1.36\,(-3.13)$	
$\eta_{\mathrm{Age<40}}$		0.170 (0.38)	0.297（0.75）	
$\eta_{\mathrm{Income<80k}}$		$-0.222\,(-0.51)$	0.984 (2.75)	
η_{PostGrad}		0.408 (0.81)	0.0849 (0.17)	
类别占比		21.59%	49.10%	29.31%
LL	-962.91	-803.51		
AIC	1,937.82	1,667.02		
BIC	1,970.31	1,829.46		

敏感度参数。相反，在不同情境下一贯选择不同路线，反映的是情境特定的利用，其驱动因素是偏差参数的变化，而不是RL意义上的系统性利用。最后，学习率较低（$\alpha_k = 0.277$）但具有显著性，这表明这些参与者学习速度较慢，会缓慢更新预期，同时依赖较早的经历来作出选择。

第*2*类 - 具有持续性剥削倾向的*DS*优先参与者：第二类约占参与者的近一半（49.10%），主要包括低收入男性参与者，他们首先体验了驾驶模拟器任务，这一点可由与这些变量相关的统计显著系数看出。该类别的参与者在驾驶模拟器中对不可靠路线表现出强烈且高度显著的偏好（$\gamma_{DS,\ Rel,\ k} = -1.49$），其幅度超过第1类的两倍。较小且显著的正向SP偏移参数（$\gamma_{SP,\ Rel,\ k} = 0.390$）表明，他们在假设任务中对可靠备选项的厌恶略低，但偏好并未发生根本变化。与第1类不同，两个敏感度参数（$\beta_{DS,\ k} = 0.337$和$\beta_{SP,\ k} = 0.837$）均具有显著性，表明其存在利用性行为，尤其是在假设情境中。学习率处于中等水平（$\alpha = 0.355$），这意味着他们适应反馈的速度略快于第1类参与者，不过在做出决策时仍在很大程度上依赖过往经历。

第*3*类——具有依赖情境的偏好和探索倾向的参与者：样本中约有29%属于此类别，其中女性和高收入参与者的比例较高。

与另外两类不同，第3类成员在驾驶模拟器中显著偏好可靠路线（$\gamma_{DS,\ Rel,\ k}$= 0.372），体现出明显的可靠性寻求倾向。然而，显著为负的SP偏移参数（$\gamma_{SP,\ Rel,\ k}$ =-0.684）表明，在假设情境中，参与者更倾向于选择不可靠路线（$\gamma_{SP,\ Rel,\ k}$ = -0.684）。两个敏感度参数均显著，但数值相对较低（$\beta_{DS,\ k}$ = 0.247和$\beta_{SP,\ k}$ = 0.141），表明与另外两类相比，他们的探索倾向更强。三类中，该类别的学习率最高（α_k = 0.437），但在统计上不显著，这表明其预期稳定，并随时间逐渐变化。最后，该类别参与者对不可靠路线的初始感知出行时间平均值为（$Q_{Unr,\ 0,\ k}$= 4.81）；相比之下，第1类和第2类参与者最初认为不可靠路线的出行时间相对较长，分别为6.16和6.12分钟，接近7分钟的最大可能出行时间。

图3 展示了三个潜在类别各自的后验类别归属概率。这些分布呈现出明显的双峰特征，大多数个体在每个类别上的概率都接近0或1。这种模式表明后验确定性很高，因为大多数个体都以较高概率归属于某一类别。这种清晰的区分证实，LCRL模型识别出了定义明确且行为特征各异的潜在类别。

此外，图4 展示了不可靠备选项的旅行时间预期、该备选项的选择概率，以及三名个体实际呈现的选择轨迹。这三名个体在三个潜在类别中的某一个类别上的后验类别归属概率均大于0.9，因此可以认为他们确定地属于各自对应的类别。这些轨迹直观呈现了表4 中的类别特定参数如何在20种决策情境中转化为行为模式。

属于第1类的个体在所有陈述偏好情景中均选择可靠路线，选择不可靠路线的概率接近于零。这一行为与较大且显著为正的陈述偏好偏移参数（$\gamma_{SP,\ Rel,\ k}$ = 3.70）一致；该参数在假设情境中强烈偏向可靠路线。切换到驾驶模拟器实验后，这种偏好发生逆转：选择不可靠路线的概率急剧上升，且无论经历何种反馈，该个体在所有驾驶模拟器情景中都选择这一备选项。这与负向且显著的驾驶模拟器偏差参数（$\gamma_{DS,\ Rel,\ k}$ =-0.635）相符，表明当驾驶模拟器中实际呈现结果时，个体明显偏好不可靠路线。在驾驶模拟任务中，预期仅逐渐调整，这反映了该类别估计出的学习率较低。

代表第2类的个体呈现出不同的模式。在驾驶模拟器和陈述偏好情景中，该参与者在所有情景中都选择不可靠路线，只有一次例外（情景3），且选择概率始终较高，保持在0.7以上。这种持续的偏好与强烈的负向DS偏差（$\gamma_{DS,\ Rel,\ k}$ =-1.49）以及幅度较小但为正的SP偏移参数（$\gamma_{SP,\ Rel,\ k}$ = 0.390）相符；这两者总体上都倾向于不可靠路线。对不可靠路线的这种利用性行为，也与较高且显著的敏感度参数值（$\beta_{DS,\ k}$ = 0.337和$\beta_{SP,\ k}$ = 0.837）一致。与第1类相比，其旅行时间预期波动更加剧烈，这与适中且显著的学习率（α_k = 0.335）相符，该学习率使其能够根据实际反馈更快地进行调整。

相比之下，归入第3类的个体表现出更多探索性行为，在两种情境中都会在两个备选项之间切换。这与相关的无码高清

较低的敏感度参数（$\beta_{DS,\ k} = 0.247$ 和 $\beta_{SP,\ k} = 0.141$）会导致更具探索性的选择模式。不过，在DS情景中，该个体表现出偏好可靠路线的倾向（10次中选择该路线6次），这与正向的DS偏差（$\gamma_{DS,\ Rel,\ k} = 0.372$）相一致。在SP情境中，这种偏好转向不可靠路线（10次中选择该路线6次），与负向且显著的SP偏移参数（$\gamma_{DS,\ Rel,\ k}$ =-0.684）相符。

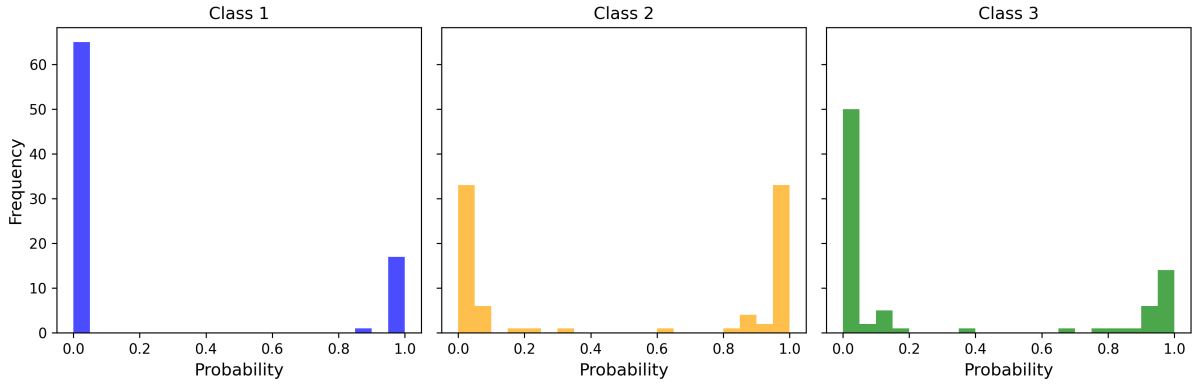


图3：类别归属的后验概率（类别成员概率）

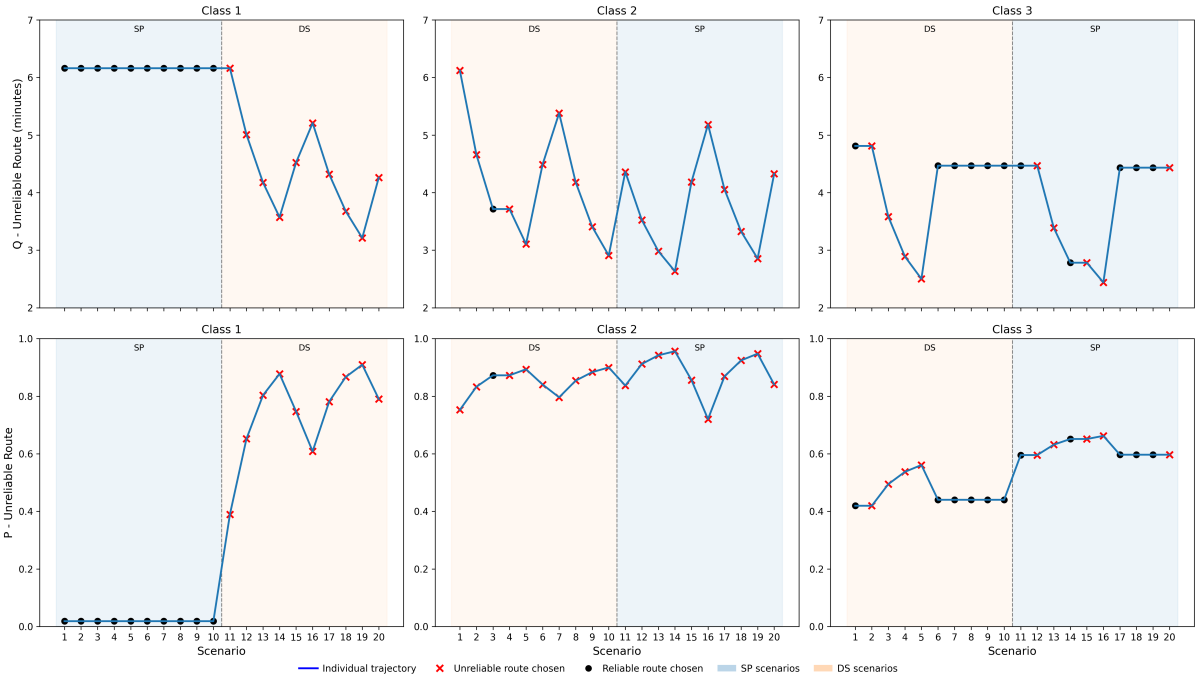


图4：不可靠路线的旅行时间预期和选择概率，以及三位个人在20种情景中的观测选择轨迹；每位个人分别对应一个不同的类别（其中10种情景来自驾驶模拟器，另外10种来自陈述偏好调查）

接下来，图5 展示了三个“极端”实验中，每个类别一位代表性个体的出行时间预期和选择概率的模拟轨迹。实验对象为不

可靠路线。在所有实验中，假设个体在20种驾驶模拟器情境中始终选择不可靠路线，同时接收以下反馈模式之一：a）恒定出行时间为2分钟；b）恒定出行时间为7分钟；或（c）按2、2、7、7、7分钟的序列重复。这些模拟情形突显了类别特定的学习率和敏感度参数如何影响持续（始终为2分钟或始终为7分钟）或交替反馈条件下预期和选择概率的演变与收敛。与表4中报告的估计结果一致，在稳定反馈下，第3类的预期收敛最快，这反映了其相对较高的学习率。在交替反馈情境下，三个类别呈现出不同的重复模式，其中第3类的预期更新最强且最快，这同样与其相较于另外两个类别较高的学习率一致。就选择概率而言，第2类在所有反馈情境下始终表现出较高的数值，反映了其对不可靠路线的强烈偏好。此外，无论收到何种反馈，第2类的选择概率都仅在0.8附近略有波动，表明其具有利用倾向。相比之下，第3类的选择概率在0.4附近波动，反映了其在驾驶模拟器情境中更倾向于寻求可靠性。

同样，图6 展示了 SP 条件下相同的模拟实验。第1类在 DS 情境下平均表现出较高的不可靠路线选择概率（图5），而图6 显示，无论收到何种反馈，该类别在所有 SP 情景下选择不可靠路线的概率都较低。这种反转反映了该群体依情境而变的偏好（在 DS 情境下强烈偏好不可靠选项，而在假设性的 SP 情境下则偏好可靠选项）。第3类的选择概率变化模式与 DS 条件下观察到的相似，但整体上移，集中在约 0.6 的较高水平（DS 条件下约为 0.4）。这种上移反映了该类别在假设性 SP 情境下更倾向于寻求风险。

总体而言，这些模拟轨迹凸显了 LCRL模型揭示的行为差异，并进一步说明了各类别如何应对极端学习环境中持续或波动的经历。

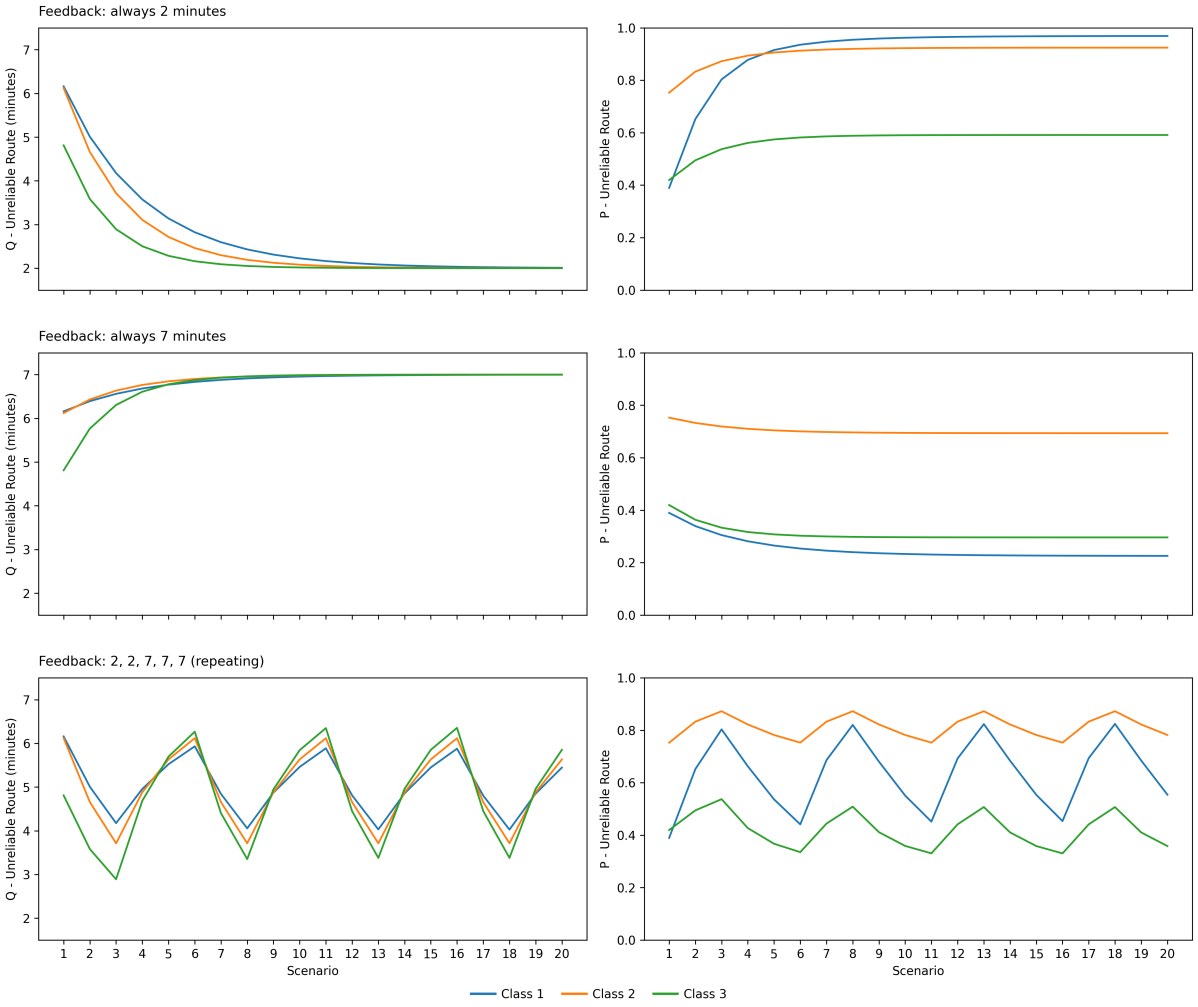


图5：各类别的期望和概率轨迹——DS情景

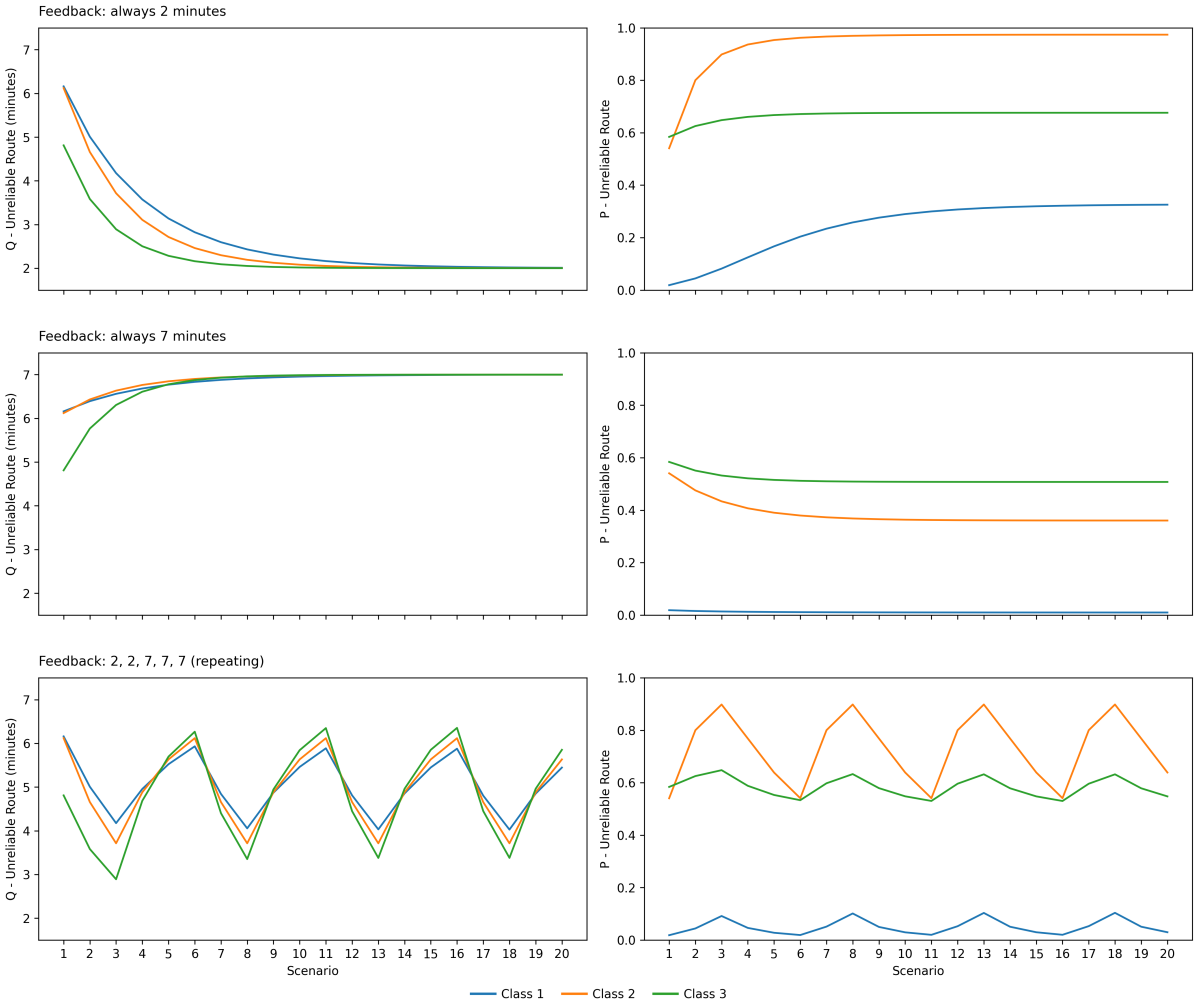


图6：各类别的期望与概率轨迹——陈述偏好情景

5 结论

本研究通过引入一种完全贝叶斯潜在类别强化学习（LCRL）模型，搭建了计量经济学离散选择模型与数学心理学之间的桥梁，并采用机器学习中的变分推断技术对模型进行估计。与传统的潜在类别选择模型（LCCM）类似，所提出的LCRL模型由两个子部分构成：潜在类别成员资格模型和类别特定选择模型。前者将未观测异质性视为离散潜在构念或类别，从而能够表示不同的学习类型，而不必假设只有一种行为过程。这种结构使模型能够揭示不同群体在探索或利用备选项、更新信念和做出选择方面的系统性差异，而这些差异通常会被标准强化学习模型或连续随机参数模型忽略。后者则利用Rescorla-Wagner强化学习模型捕捉类别特定的学习动态。

传统离散选择模型通常用于捕捉静态条件下的稳定偏好，而强化学习模型则明确关注通过过往经历所获得的反馈或奖励不断演化的动态适应性行为。

通过整合这两种视角，所提出的LCRL模型能够更全面地理解个人如何随时间动态更新其决策偏好。具体而言，强化学习模型通过潜在类别划分获得了考虑未观测异质性的能力，从而捕捉不同个人之间的学习动态。反过来，计量经济学中的潜在类别离散选择模型则受益于通过类别特定强化学习选择模型对动态偏好演化以及探索-利用权衡行为进行建模的能力。

LCRL模型在一个通过驾驶模拟器和陈述偏好实验收集的路径选择数据集上进行了检验。其表现显著优于基准强化学习模型，并揭示了三种具有不同学习策略的潜在类别。实证结果显示：(i) 一类具有依情境而变的偏好者，主要由SP优先参与者组成；他们在假设情境中会选择可靠路线，但在驾驶模拟器中亲身体验结果时，则转而选择不可靠路线；(ii) 一类利用倾向很强的参与者，他们在DS和SP情境中始终偏好不可靠路线，并能适度地根据反馈进行调整；(iii) 一类探索倾向更强的参与者，其中女性和高收入参与者的可能性更高；他们在驾驶模拟任务中偏好可靠路线，但在SP条件下则会反转这一偏好。这些发现表明，个体之间的差异不仅体现在偏好上，也体现在学习方式、更新预期的速度以及平衡探索与利用的方式上。这些发现对旅行行为建模也具有重要意义，因为它们有助于解释陈述选择与实际体验选择之间的差异，并能更深入地理解出行者如何适应交通网络中的变异性与不确定性。

除了本研究带来的出行行为洞见之外，方法本身的创新也是一种新兴建模技术。将计量经济学选择模型与强化学习和贝叶斯推断相结合，为捕捉适应性、基于经验的决策提供了一个新框架。虽然我们的实证应用聚焦于人类路线选择，但这种方法与新兴出行技术直接相关。例如，该方法可为自动驾驶汽车（AV）的路径规划策略开发提供参考。由于自动驾驶汽车将在由人类决策者塑造的环境中运行，所提出的LCRL模型可支持设计自动驾驶汽车路径规划算法，使其更好地预测人类的学习和预期形成过程（Zhang 等人，2025），最终在混合自动化系统中促进符合社会利益且为用户所接受的导航策略。同样，所提出的LCRL模型也适用于其他以学习、适应和异质性为核心的新兴出行场景。在出行即服务（MaaS）平台上，出行者常会尝试不熟悉的出行方式或捆绑选项，之后才逐渐形成习惯性的使用模式。基于LCRL的方法可补充现有依赖静态潜在类别聚类的MaaS细分研究（Gan 和 Li，2025），将动态学习行为纳入分析，并识别哪些群体更可能进行探索，以及激励或信息如何加快采纳进程。同样，实时信息和个性化路径引导系统越来越依赖行为预测模块来塑造用户互动。LCRL框架能够同时捕捉预期更新以及类别特定的探索—利用倾向，因此有望成为设计信息策略的有效工具，使策略能够识别适应性行为，而非假定偏好一成不变。对于出行者必须偏离习惯行为的情形，该模型也能为交通规划和政策制定提供宝贵洞见。由于它能够捕捉个人如何探索新选项，以及在意外结果后如何更新预期，

该模型可帮助政策制定者预判出行者对网络关闭、突发事件或服务不可靠等干扰的行为反应。同样，该模型还可用于评估出行者如何学习使用新建基础设施，例如新开通的地铁线路、高速公路连接线或自行车道，并识别哪些群体更倾向于尽早尝试，以及不同群体需要多久才能形成新的稳定模式。因此，纳入这些适应性学习过程，有助于在网络变化期间实现更准确的需求预测，并制定有针对性的信息或激励策略。

最后，未来的研究也可以通过允许强化学习参数因个体而异，来探讨类别内异质性。

致谢

作者感谢欧洲研究委员会通过高级研究资助 101020940-SYNERGY，以及澳大利亚研究委员会通过探索项目 DP150103299 提供的资金支持。

附录 A

标准强化学习（RL）模型的生成过程概述如下：

1. 抽取强化学习探索—利用敏感度参数 $\beta \sim \text{Lognormal}(\mu_\beta, \sigma_\beta)$，确保 $\beta > 0$。

2. 抽取强化学习学习率参数 $z_\alpha \sim \mathcal{N}(\mu_{z_\alpha}, \sigma_{z_\alpha})$，然后应用 Sigmoid 变换，确保学习率位于 $[0,1]$：$\alpha_0 = S(z_{\alpha_0}) = \frac{1}{1+e^{-z_{\alpha_0}}}$

3. **对于每个备选项** $i \in \{1, \ldots, I\}$:

(a) 抽取强化学习偏差参数 $\gamma_i \sim \mathcal{N}(\mu_{\gamma_i}, \sigma_{\gamma_i})$（出于识别目的，每个类别中有一个备选项的参数设为零。）(b) 抽取初始期望 $z_{Q_{i0}} \sim \mathcal{N}(\mu_{z_{Q_{i0}}}, \sigma_{z_{Q_{i0}}})$，然后应用以下 Sigmoid 变换，确保 $[a,b]$ 中的初始 Q值为正： $Q_{i0} = a + (b-a) \times S(z_{Q_{i0}})$，其中 a 和 b 由实验设计固定，不作估计。

4. **对于每个个体** $n \in \{1, \ldots, N\}$:

(a) **对于每个选择时点** $t \in \{1, \ldots, T_n\}$:

i. 抽取观测选择 $y_{nt} \sim$ 多项Logit$(\beta, \{\gamma_i, Q_{nit}\}_{i=1}^I)$, ii. 计算所选备选项的预测误差 $y_{nt} = j$ （根据步骤 i）： iii. 更新所选备选项的期望 $y_{nt} = j$ （根据步骤 i）：

$$\delta_{njt} = r_{njt} - Q_{njt}$$

$$Q_{nj(t+1)} = Q_{njt} + \alpha\, \delta_{njt}$$

附录B

本附录介绍一项参数恢复练习，用于验证所提出的 LCRL模型及其变分推断程序能否可靠地恢复生成数据的底层真实参数。参数恢复是一种基于模拟的验证技术，具体做法是使用已知“真实”参数从模型生成合成数据集，在这些生成的数据集上估计模型，然后将恢复估计值与原始值进行比较。成功恢复参数可为以下方面提供证据：模型具有可识别性，推断程序实现正确，认知假设应用于真实数据时成立，并且对参数估计值的认知与行为解释有效，而非估计过程或模型设定造成的假象。

在本练习中，我们生成100个合成数据集，以复现第3节真实数据集的结构和实验设计，以及第4节引入的适应过程。每个模拟数据集均采用相同数量的受访者（83名）、每位受访者相同的社会经济和人口统计特征、每位受访者相同数量的试次（20次）、相同的实验设置（10次驾驶模拟器选择和10次陈述偏好选择），以及相同的选择情境：在5分钟的可靠路线和一条耗时可能介于2至7分钟之间的不可靠路线之间进行选择。使用双类别LCRL模型，我们为每次模拟抽取一组新的类别特定参数：可靠选项的基准偏差参数（$\gamma_{DS、Rel、k}$）来自均值为0、标准差为1的正态分布；SP偏移偏差（$\gamma_{SP、Rel、k}$）来自同一分布；DS和SP情境下的探索—利用敏感度（$\beta_{DS、k}$ 和$\beta_{SP、k}$）来自0.1至2之间的均匀分布；学习率（α_k）来自0.05至0.095之间的均匀分布；不可靠路线的初始预期时间（$Q_{Unr、0、k}$）来自2至7之间的均匀分布；类别归属参数（η_k）来自均值为0、标准差为1的正态分布。对于每位受访者，双类别LCRL模型使用这组新参数生成完整的选择与反馈序列。由此得到100个合成面板数据集，这些数据集由已知的“真实”参数生成，但在结构上与真实数据相似。随后，使用与第4 节实际应用中相同的LCRL设定和推断设置（例如变分推断、先验和约束）对这些模拟数据集进行拟合。

图7展示了 100 个模拟数据集中每个估计参数的后验均值与其对应真实值之间的关系。图中结果表明，所提出的 LCRL 模型能够有效恢复生成数据的潜在参数。类别特定强化学习选择参数紧密分布在 45度线附近，仅有轻微离散。尽管在较短的重复选择序列中，学习率通常较难恢复，但其估计值仍与模拟值高度吻合。此外，特定情境下的偏差和敏感度参数也能得到较好的恢复，这支持了模型设定中的 DS-SP分离。同样，不可靠路线的初始预期时间几乎得到了完美恢复，表明模型成功区分了先验信念与后续更新。类别归属参数将社会经济和人口特征与潜在类别联系起来，其在 45度线附近的变化略大。这是预期之中的，因为这些参数的可识别性取决于每次模拟中各类别之间的行为距离，而某些模拟参数集可能会产生相对相似的类别。尽管如此，图中结果总体表明参数恢复较为可靠。除定性目视检查外

, 我们计算定量指标，以

概括各参数在所有模拟中的恢复效果。对于每个参数和潜在类别，我们计算平均偏差（即估计值与真实值之差的均值）、归一化均方根误差（NRMSE）、真实值与恢复值之间的相关性，以及决定系数（R^2）。所有指标均根据估计参数的后验均值及其对应的真实值计算。

归一化均方根误差通过 100 个模拟数据集中真实参数值的取值范围对均方根误差进行缩放，从而能够比较不同尺度的参数：

$$\text{NMRSE} = \frac{\sqrt{\frac{1}{S}\sum_{s=1}^{S}(\hat{\theta}_s - \theta_s)^2}}{\max(\theta_s) - \min(\theta_s)}, \tag{19}$$

ˆ 其中， S 为模拟数据集的数量， θ_s 为真实参数， θ_s为相应恢复的后验均值。

决定系数定义如下：

$$R^2 = 1 - \frac{\sum_{s=1}^{S}(\hat{\theta}_s - \theta_s)^2}{\sum_{s=1}^{S}(\theta_s - \text{mean}(\theta_s))^2} \tag{20}$$

表 5中汇总的恢复统计结果证实，所提出的 LCRL模型能够在 100 个模拟数据集中可靠地恢复底层数据生成参数。所有参数的偏差值都接近于零，归一化均方根误差值通常较低，表明估计误差相对于真实参数变化范围较小。真实估计值与恢复估计值之间的相关性通常超过 0.80，且 R^2 大多数参数的值都高于 0.75，这表明变分推断过程能够稳定地重建原始参数空间的结构。

值得注意的是，类别成员资格参数的变异性最大，相关性介于 0.68 和 0.85 之间，归一化均方根误差值在 0.10–0.15 之间，且 R^2 值介于 0.45 和 0.72 之间。然而，如前所述，这些参数的可识别性取决于每次模拟中各类别之间的行为差异，某些模拟抽样可能会产生相似或重叠的行为类别。不过，这些数值仍表明此类参数能够得到中等程度但可靠的恢复。

总体而言，恢复结果有力地表明，所提出的LCRL模型具有良好的可识别性，能够可靠地恢复影响学习动态和潜在类别划分的参数，从而支持第4节报告的实证结果及我们据此作出的行为解释。

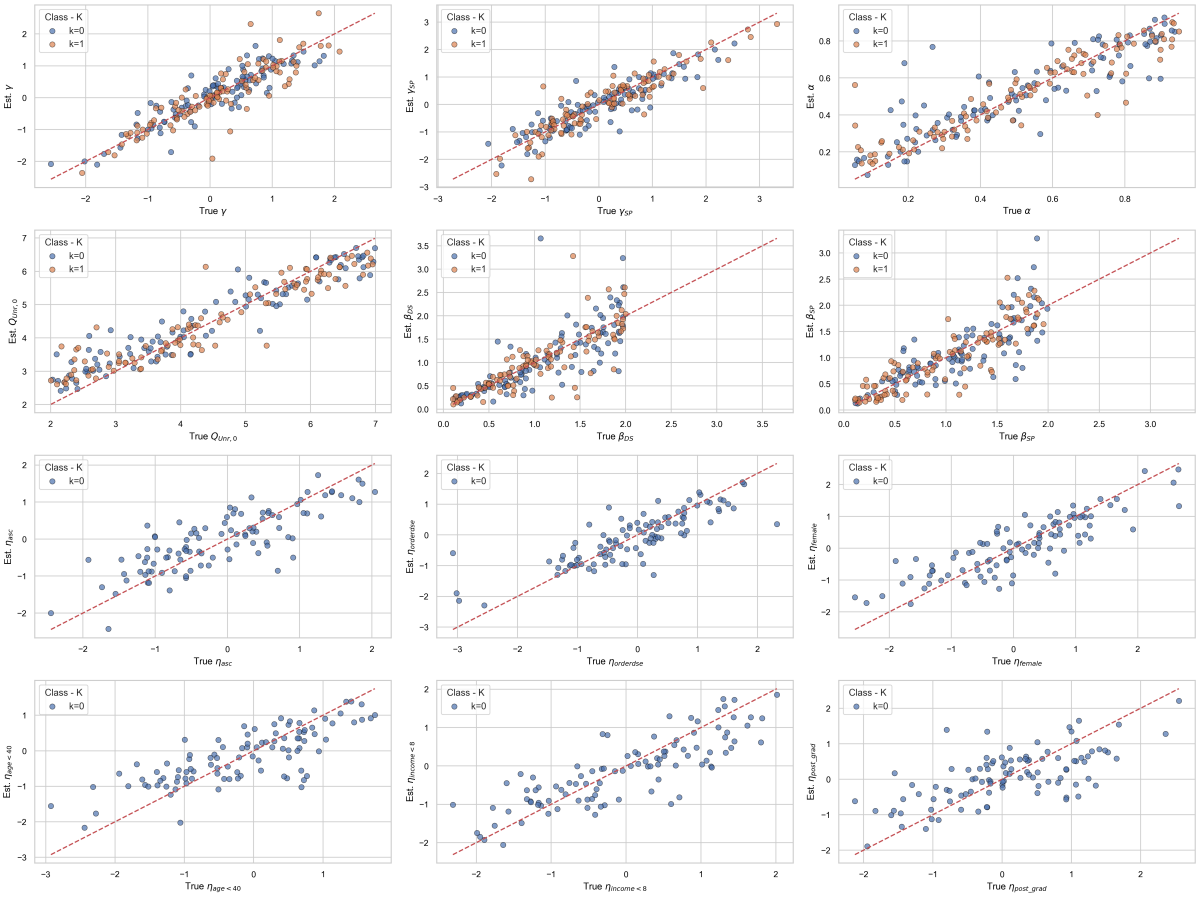


图7：LCRL模型在模拟数据集上含两个类别的参数恢复图

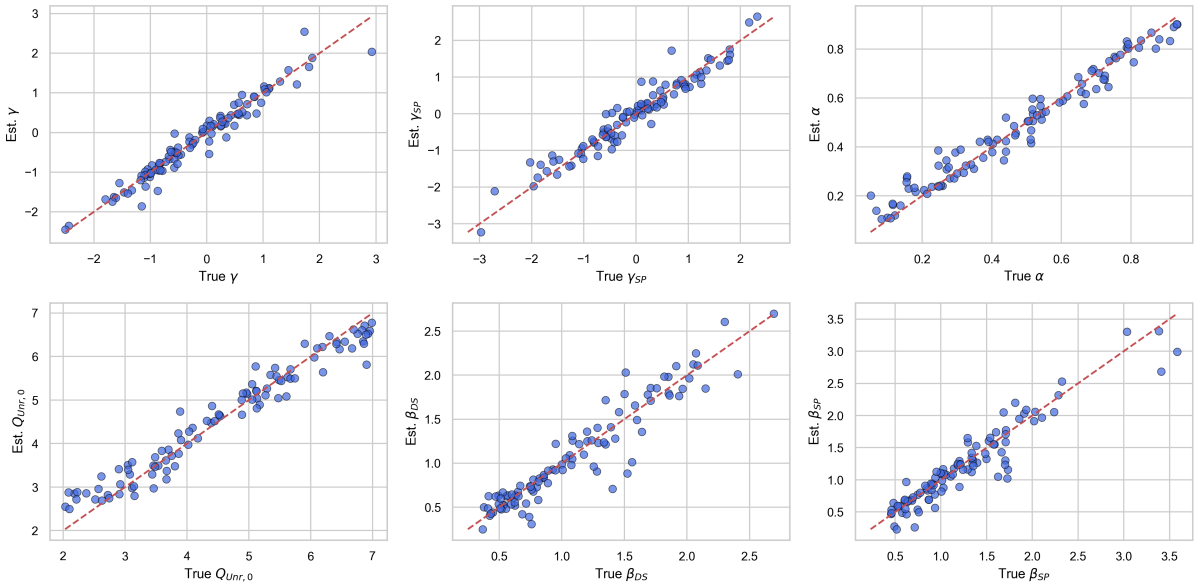


图8：RL模型在模拟数据集上的参数恢复图

表5：LCRL模型在模拟数据集上的参数恢复指标

参数 k		偏差	归一化均方根误差	相关性	R^2
$\gamma_{\text{DS,Rel,k}}$	0	0.018	0.086	0.901	0.804
$\gamma_{\text{DS,Rel,k}}$	1	-0.038	0.107	0.883	0.746
$\gamma_{\text{Rel,SP,k}}$	0	-0.062	0.096	0.893	0.790
$\gamma_{\text{Rel,SP,k}}$	1	-0.060	0.083	0.916	0.828
α_{k}	0	0.014	0.141	0.880	0.772
α_{k}	1	0.002	0.117	0.919	0.837
$Q_{\text{Unr,0,k}}$	0	0.120	0.105	0.949	0.882
$Q_{\text{Unr,0,k}}$	1	0.031	0.108	0.941	0.870
$\beta_{\text{DS,k}}$	0	-0.019	0.238	0.781	0.420
$\beta_{\text{DS,k}}$	1	-0.019	0.193	0.819	0.565
$\beta_{\text{SP,k}}$	0	-0.025	0.194	0.790	0.490
$\beta_{\text{SP,k}}$	1	0.003	0.155	0.886	0.742
η_{ASC}	0	0.073	0.125	0.815	0.657
$\eta_{\text{DS 第一}}$	0	-0.021	0.105	0.823	0.677
η_{Female}	0	-0.006	0.109	0.848	0.719
$\eta_{\text{Age<40}}$	0	0.014	0.135	0.781	0.609
$\eta_{\text{Income<80k}}$	0	-0.092	0.137	0.832	0.685
η_{PostGrad}	0	0.016	0.150	0.678	0.453

表6：强化学习模型在模拟数据集上的参数恢复指标

参数	Bias	归一化均方根误差	相关性	R^2
γ_{DS}	-0.024	0.041	0.973	0.946
γ_{SP}	0.023	0.052	0.965	0.930
α	0.006	0.057	0.981	0.958
$Q_{\text{Unr,0}}$	0.032	0.068	0.977	0.946
β_{DS}	-0.020	0.082	0.944	0.880
β_{SP}	-0.042	0.069	0.944	0.886

此外，我们还使用与LCRL模型相同的模拟，对基准强化学习模型进行参数恢复分析，结果见第4节。图8显示，所有强化学习参数均得到了近乎完美的恢复，因为各值都紧密聚集在45度线周围。同样，表6 表明，所有参数的恢复效果都极为出色。偏差值接近于零，归一化均方根误差值也始终很低（小于0.1），这表明相对于模拟参数的尺度，估计误差极小。所有参数的相关性均超过0.94，且 R^2 值均高于0.88，说明恢复估计值与真实的数据生成值高度吻合。

参考文献

Cantillo, V., de Dios Ort´uzar, J., & Williams, H. C. W. L. (2007年). 惯性和序列相关存在时的离散选择建模. 《*Transportation Science*》, *41*(2), 195-205. doi: 10.1287/trsc.1060.0178

Cherchi, E., & Manca, F. (2011年). 交通方式选择中的惯性：基于RP/SP数据集的新证据。《*Transportation*》， *38*(4), 679-695. doi: 10.1007/s11116-011-9338-9

Cirillo, C., & Xu, R. (2011年). 交通领域的动态离散选择模型。《*Transport Reviews*》， *31*(4), 473–494. doi: 10.1080/01441647.2010.533393

Collins, A. G. E., & Shenhav, A. (2022). 神经科学中学习与决策建模的进展。《*Neuropsychopharmacology*》， *47*(1), 104-118. doi: 10.1038/s41386-021-01126-y

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Kasper, L., Mauer, J., & Stephan, K. E. (2017, 01). 社会学习过程中中脑和隔区的层级预测误差。《*Social Cognitive and Affective Neuroscience*》， *12*(4), 618-634. doi: 10.1093/scan/nsw171

Fayyaz, M.（2021年）。风险与不确定出行时间下驾驶员的路线选择行为——陈述选择与模拟体验（博士论文，悉尼大学商学院交通与物流研究所，澳大利亚悉尼）。取自 https://hdl.handle.net/2123/26723

Gan, Z., & Li, W.（2025）。谁是MaaS回避者、观望者或爱好者？是什么推动他们采用MaaS的意愿？《*Transport Economics and Management*》，*3*，57-69。doi: https://doi.org/10.1016/j.team.2025.02.001

Gao, K., Yang, Y., Sun, L., & Qu, X.（2020）。揭示交通方式转变行为中的心理惯性及其对通勤出行的量化影响。《*Transportation Research Part F: Traffic Psychology and Behaviour*》，*71*，272-287。doi：10.1016/j.trf.2020.04.006

Henriquez-Jara, B., & Guevara, C. A. (2025). 基于经验的选择模型（EBCM）：构式、识别、行为洞见与福祉评估。《*Journal of Choice Modelling*》，*55*，100552。doi: https://doi.org/10.1016/j.jocm.2025.100552

Henriquez-Jara, B., Guevara, C. A., Munizaga, M., & Perez, O. D. (2025). 习惯与更优交通选项的次级探索：一种双系统方法。《旅行行为与社会》，*38*，100877。doi: https://doi.org/10.1016/j.tbs.2024.100877

Hess, S. (2024). 潜在类别结构：偏好异质性及其他方面。载于 S. Hess & A. Daly（编），《选择建模手册》（第 372–391 页）。Edward Elgar。doi: 10.4337/9781800375635 Hess, S., & Train, K. E. (2011). 使用混合Logit模型恢复个体间与个体内异质性。*Transportation Research Part B: Methodological*，*45*（7），973-990。doi: 10.1016/j.trb.2011.05.002 Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). 随机变分推断。《机器学习研究杂志》，*14*，1303 – 1347。Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). 图模型变分方法简介。《机器学习》，*37*（2），183–233。doi: 10.1023/ A:1007665907178

Kingma, D. P., & Ba, J. (2017). Adam：一种随机优化方法。*arXiv* 预印本 *arXiv:1412.6980*。取自 https://arxiv.org/abs/1412.6980

Lei, W.、Liu, K.、Chen, G.、Tolomeo, S.、Liu, C.、Peng, Z.、.....Chen, J.（2022）。情感唤醒对人类基于奖励的决策中的选择行为、奖励预测误差和反馈相关负波的影响。*Frontiers in Psychology*, *52*（11）, 2124–2133。doi：10.1017/S003329172000402X

Liu, H.-H.、Hsieh, M. H.、Hsu, Y.-F.和Lai, W.-S.（2015）。情感唤醒对人类基于奖励的决策中的选择行为、奖励预测误差和反馈相关负波的影响。*Frontiers in Psychology*, *6*。doi：10.3389/fpsyg.2015.00592

Niv, Y.（2009）。大脑中的强化学习。*Journal of Mathematical Psychology*, *53*（3）, 139-154。doi：10.1016/j.jmp.2008.12.005

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). 人类大脑中的时序差分模型与奖励相关学习。*Neuron*, *38*(2), 329-337. doi: 10.1016/S0896-6273(03)00169-7

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., . . . Chintala, S. (2019). PyTorch：一款采用命令式风格的高性能深度学习库。载于 H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alch´e-Buc, E. Fox, & R. Garnett (编), 神经信息处理系统进展 (第 32 卷)。Curran Associates, Inc. Rad, J. A., Hess, S., Hancock, T. O., Bliemer, M., Beck, M., Khan, F., & Ben-Elia, E. (2025). 这条路通向何方？强化学习模型对路线选择行为的启示。*PsyArXiv*. doi: 10.31234/osf.io/25sc9 v1 Rescorla, R., & Wagner, A. (1972). 巴甫洛夫条件作用理论：强化与非强化效力的变化。Appleton-Century-Crofts. Rodrigues, F. (2022). 将混合多项式Logit模型的贝叶斯推断扩展至大规模数据集。*Transportation Research Part B: Methodological*, *158*, 1-17. doi: https://doi.org/ 10.1016/j.trb.2022.01.005 Samson, R. D., Frank, M. J., & Fellous, J.-M. (2010). 强化学习计算模型：多巴胺作为奖励信号的作用。*Cognitive Neurodynamics*, *4*(2), 91-105. doi: 10.1007/s11571-010-9109-x Schultz, W., Dayan, P., & Montague, P. R. (1997). 预测与奖励的神经基础。*Science*, *275*(5306), 1593-1599. doi: 10.1126/science.275.5306.1593 Steingroever, H., Wetzels, R., & Wagenmakers, E.-J. (2014). 强化学习模型在爱荷华赌博任务中的绝对表现。*Decision*, *1*(3), 161–183. doi: 10.1037/dec0000005 Sutton, R. S., & Barto, A. G. (2018). 强化学习：导论。第2版，MIT Press. V´elez, N., & Gweon, H. (2022). 向他者心智学习：对社会学习强化学习模型的乐观批判。*Current Opinion in Behavioral Sciences*, *38*, 110–115. doi: 10.1016/j.cobeha.2021.01.006 Wainwright, M. J., & Jordan, M. I. (2008). 图模型、指数族与变分推断。*Foundations and Trends Machine Learning*, *1*(1-2), 1–305. doi: 10 . 1561/2200000001

Zhang, X., Hess, S., Hancock, T., Zhong, S., Jia, N., &Zhang, Q. (2025). 出行方式选择中的协调行为：一项关于自动驾驶汽车服务的实验。 取自 http://dx.doi.org/10.2139/ssrn.5207101 (SSRN 预印本 5207101)