INTRODUCING DUAL-SYSTEM THEORY TO TRAVEL BEHAVIOR: THE RELATIONSHIP BETWEEN HABITS AND SUB-EXPLORATION OF NOVEL AND BETTER TRANSPORTATION OPTIONS

Bastián Henríquez-Jara, University of Chile – bastian.henriquez@ug.uchile.cl
C. Angelo Guevara, University of Chile, Complex Engineering Systems Institute (ISCI) – crguevar@gmail.com
Marcela Munizaga, University of Chile, Complex Engineering Systems Institute (ISCI) – mamuniza@uchile.cl
Omar D. Perez, University of Chile, Complex Engineering Systems Institute (ISCI), California Institute of Technology – omar.perez.r@uchile.cl

Keywords: habits, reinforcement learning, travel behavior.

ABSTRACT

In this article, we introduce dual-system theory to travel behavior literature, which allows a better understanding of the non-compensatory nature of habits in commuting decisions. In a laboratory task, we show how the individual degree of habitual behavior can be estimated with Reinforcement Learning. Furthermore, we show that the exploration of new (and better) alternatives in a transportation choice scenario is hindered when subjects develop habit-like strategies. Our results suggest that sub-optimal routes in real life may be due to sub-exploratory behavior, who may benefit from policies aimed at prompting exploration of new alternatives once they are implemented.

1. INTRODUCTION

During the first week of February 2014 there was a major change in the daily routine of Londoners. After the announcement that redundancies and ticket office closures would occur, the Rail Maritime Transport union went on a 48 hour strike which resulted in more than 60% of the stations being at least partially closed (Larcom et al., 2017). The primary implication of this shock was that it compelled commuters to explore new routes they would not have otherwise chosen for their daily trips. Larcom et al. (2017) found that around 5% of commuters stuck with alternative routes even after the strike ended, i.e., for some reason they preferred the route they were forced to explore. A likely hypothesis is that habits caused sub-exploration of the available routes prior to the strike. Our work develops a possible explanation using the dual-system theory of behavior, a theory that is widely accepted in psychology and neuroscience (Perez & Dickinson, 2020; Daw & O'Doherty, 2013).

This natural experiment provide more evidence to question the canonical assumptions deployed in the Random Utility Models (RUMs, McFadden (1974)) widely used in travel behavior research.

This theory assumes that commuters make choices based on the subjective satisfaction or utility provided by each of the option's attributes (e.g. travel time, price, and crowding). More importantly, RUM assumes compensatory behavior, i.e., if any attribute of an alternative changes, the utility that this alternative gives to an individual can be recovered by properly varying the other attributes. For example, the disutility of higher prices could be compensated by a reduction in travel time or other relevant attributes. Compensatory behavior is key for welfare analysis, including the critical concept of "value of time" in transportation.

However, empirical data of transport choices (as Larcom et al. (2017)) have shown that people are neither always utility maximizers nor do they behave in a compensatory way. People may be "*rational*", but they do not act as "*global utility maximizers*" (Miller, 2020). For this reason, travel behavior theories have attempted to incorporate psychological aspects to gain a deeper understanding of the underlying processes behind real-life choices.

Most travel studies incorporating psychological aspects still appeal to strong rationality and compensatory behavior assumptions. The role of the unconscious or automaticity in decision-making has been underexplored (Vaa, 2014). When taking into consideration automaticity factors, the rationality assumptions shift to *bounded rationality* assumptions: people are rational but with limited mental capabilities.

The travel behaviour literature has made progress in bounded rationality models (Rasouli & Timmermans, 2014). However, in our view, one of the most important gaps in the travel behavior literature is how habitual behavior is conceived.

Typically, the influence of habits on choices is modeled in two different ways: based on the habits persistence model (Heckman, 1981) (basically, with lagged variables or latent Markov models), or with a latent inertia (Gao et al., 2020). However, neither of these methods satisfy the most important characteristic of habitual behavior: the lack of a conscious deliberation process.

In the cognitive psychology and neuroscience literature, it is a long-standing idea that a rewarded behavior (a behavior followed by a positively valued outcome) which is consistently performed in a similar situation renders the behavior automatic and independent of the expected utility of its outcomes or consequences, which is defined as habitual behavior (Dickinson & Balleine, 1994). Clearly, in the presence of habits, there is no compensatory behavior, which was pointed out by Gärling & Axhausen (2003), but has not yet been reflected in choice modeling.

Under the above definition of habits, it is necessary to know if decisions are deliberated or not to classify them as habitual. This paradigm, the duality of deliberated and automatic behavior, is the basis of dual-system theory. In dual-system theory, human (and animal) behavior can be broadly represented by two types of learning models, the Model-free (MF) and the Model-based (MB) (Gershman, 2015). MF behavior represents habit-like strategies, and MB behavior represents utility-based (or goal-directed, in the psychological jargon) strategies. However, both systems interact, causing the observable behavior to be a mixture of both systems.

To detect MB and MF behavior, Daw et al. (2011) developed an experimental task, known as the Markovian Two-Step Task, which is nowadays well-established on the cognitive psychology and

neuroscience literature. In this task, subjects choose between two options with stochastic reward probabilities. The responses are then modelled using reinforcement learning algorithms that allows the identification of the importance of each system in the behavior of each participant.

Using Daw et al.'s (2011) task in a transportation scenario, we test the relationship between the type of behavioral strategy deployed by subjects (MF or MB) and the degree to which those strategies are correlated with exploratory behavior in a transportation scenario.

The main contributions of this article are: (1) the introduction of the concepts of MB and MF behavior to the transportation and discrete choice literature, which allows a better understanding of habits in commuting behavior, (2) providing a modeling framework to estimate the degree of habitual (automatic) behavior for each individual, and (3) evidence that habit-like behavior is inversely related to exploration of new alternatives, which, to the best of our knowledge, is a novel result in both transportation and psychology.

The remaining of this article is structured as follows. In the second section we depict the experimental task, and in the third section we describe the computational modeling of behavior. We discuss the results in the last section of the paper. The Appendix offers a brief overview of the main RL concepts, necessary to follow the methods and discussion of this article.

2. A TRANSPORTATION MF/MB DISTINCTION EXPERIMENTAL TASK

An scheme of the experimental task is shown in 1. The main feature of this Two-step task is that it allows the experimenter to determine if subjects' choices are based on the transition probabilities between states (and thus rely on MB strategies), or whether they are simply repeating previous choices based on their past reward outcomes (indicating the use of MF or habitual strategies).

The task comprises two phases (1). The first phase is composed by two stages. In the first stage, participants must choose which of two neighbors to ask for a ride to the bus stop. Our cover story[1] for the experiment explained to participants that their neighbors did not always go to the same bus stop. For instance, neighbor A might go to the red bus more often while neighbor B would go to the blue bus more often. In the second stage, participants did not make any further choices, and the bus either arrive on time or late. They were told this uncertainty was due to congestion. In practice, this means that the reward of trial t (r_t) is positive when the bus arrives on time ($r_t = 1$) and negative when it does not ($r_t = -1$) [2]. The probability of receiving a reward was modeled by a Gaussian random walk with $SD = 0.025$ and reflecting boundaries at 1 and 0, which ensured that participants had to continually assess which option was best at any point during the task, encouraging continuous learning. The reward probability of the blue bus started at 0.5 and the probability of the red bus started at 0.3, which guarantees $E(r_t|blue\ bus) > E(r_t|red\ bus)$. It is important to mention, that participants were told they could leave the experiment at any time,

[1] The cover story was shown to participants in a video. A version in English is available at `https://youtu.be/39QevydE_oE`

[2] Fujii & Kitamura (2003) proposed a similar assumption, arguing that people frame the travel time in a dichotomous way: being later or on time.

without any consequence on the payment they were receiving.

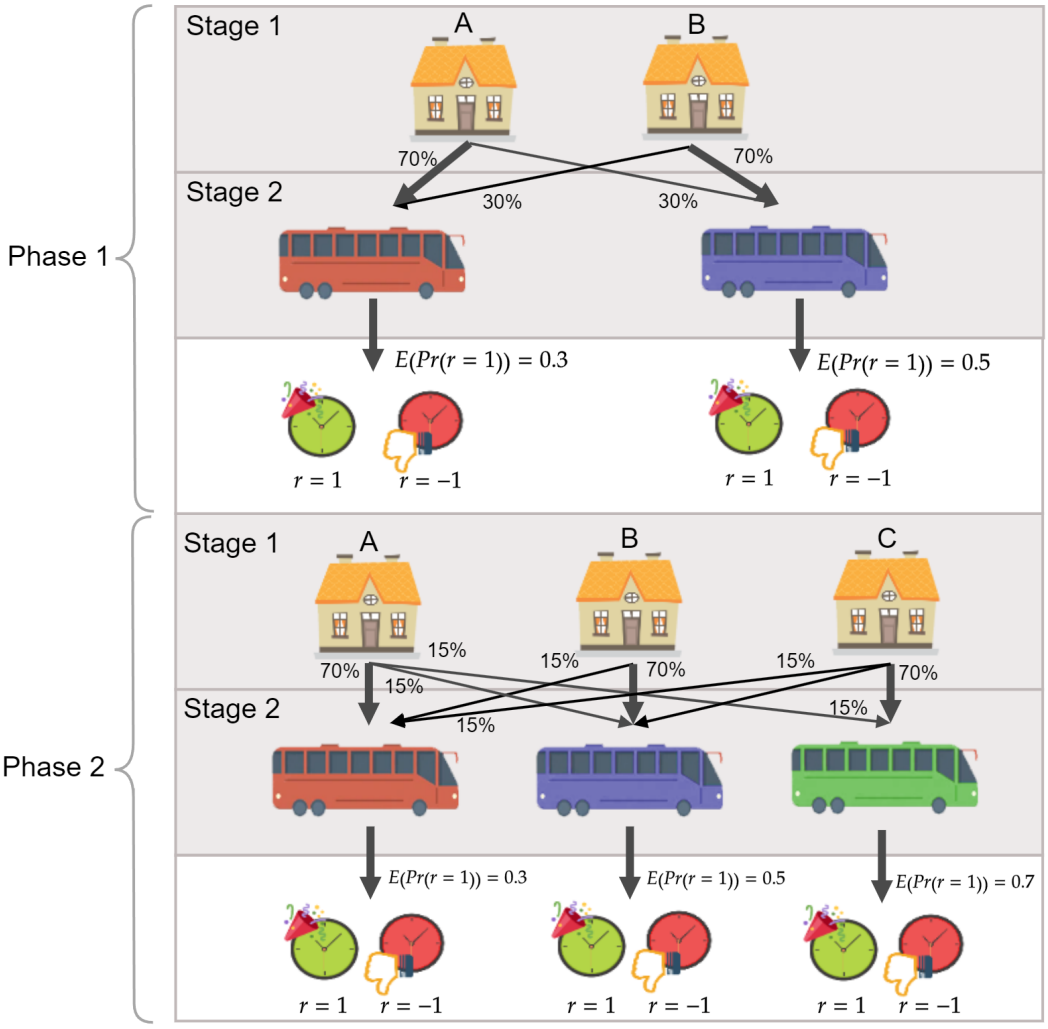

Figure 1: **A Markovian two-step task with stochastic transitions between states.** Phase 1: MB/MF distinction. At stage 1, participants make a choice between two options, in this case between two different neighbors who will take them probabilistically, to the red bus or the blue bus which, At stage 2, and without the participant having to make another choice, the bus reaches its destination with two possible outcomes: get on time (signified by a green clock) or being late (signified by a red clock). $E(r_t|blue\ bus) > E(r_t|red\ bus)$. Phase 2: exploration. A new bus was included in the choice set. The new bus had an objectively higher reward rate than the others.

The second phase aimed to measure the tendency to explore a novel alternative (Figure 1). Here, we introduced a novel option, a new bus line with a higher reward rate than the previous two options (probability of reward = 0.7), while maintaining the structure of the initial stage. Our hypothesis was that the type of behavioral control (MB/MF) used by participants would affect their exploratory behavior. Specifically, we expected that those using MF strategies would be less likely to explore new alternatives compared to MB subjects, what would reflect in an incapacity to become fully aware of the goodness of the green bus option. To model participants' behavior, we used MF and

MB RL algorithms to estimate the utility of each option on a trial-by-trial basis, and a decision rule that assigned higher choice probabilities to options with higher expected utilities. We assumed that decisions were driven by a convex combination of MF and MB strategies in each trial (Gillan et al., 2015; Nussenbaum et al., 2020).

2.1. Participants and Apparatus

Participants were called through the institutional forum of the University of Chile. 34 participants participated in this study, 15 female and 19 male, all students from the University of Chile. They were paid CLP 8.000 (around USD 8) for their participation. All participants gave their informed consent and were tested individually in cubicles in groups of a maximum of 4 subjects per session. The experiment was programmed in Psychopy (Peirce, 2007) and ran on macOS Sierra (c) desktop computers of the Transportation Laboratory at the Civil Engineering Department of the University of Chile. The experiment received IRB approval from the Nuffield College Centre for Experimental and Social Sciences at the University Santiago of Chile.

3. COMPUTATIONAL MODELING AND DATA ANALYSIS

To model participants' behavior, we used a particular kind of RL algorithm called Q-learning (Watkins & Dayan, 1992). These kinds of models are assumed to behave within Markovian chains, where the error of the estimation in t is assumed to not be transferred to time $t+1$. For the sake of simplicity, we will omit the error term across all the model formulation.

As explained before, we constructed a Markovian model of two stages $S = \{S_1, S_2\}$. The first stage (S_1) contains a set $\mathcal{C}$ with two alternatives, i.e. the two neighbors ($\mathcal{C} = \{A, B\}$). The second stage contains no alternatives to choose, but has two possible states, the red and the blue bus states ($S_2 = \{s_{blue}, s_{red}\}$). We denote each trial of the experiment as $t \in \{1, ..., T\}$, where T is the total number of trials.

We will denote the choice in the first stage (neighbor choosing stage) by $y_{i,t} = 1$, if the subject chooses i in trial t, and $y_{i,t} = 0$ otherwise. The computational modeling is based on the assumption that, in each trial t, subjects associate a value to each alternative j ($QNET_{t,j}$), which is a mixture of the values that each cognitive system associates to each alternative ($QMB_{t,j}$ and $QMF_{t,j}$). We call it "value" to adopt neuroscientist's terminology, but for general purposes it can be considered a utility estimated as a convex combination of two utilities proceeding from different models. As we will later discuss, the behavioral combination could be also captured with a latent class approach, however in this article we adopted how this is modeled in human RL literature. The estimation of $QNET$, QMF and QMB is detailed in the next subsection.

The RL model is fitted separately for each phase of the experiment and separately for each subject. In the first phase, we were interested in estimating for each participant which system weighs more on their decision (MB or MF), while in the second test phase we were interested in estimating

which participants explored more the novel alternative.

QMF estimation (Model Free algorithm)

On each trial t, the MF system associates a value $QMF_{t,j}$ to each alternative j, ignoring the transition probabilities, which is the critical assumption of MF system and represents the automaticity of behavior. $QMF_{t,j}$ is a learned value that is updated in each trial based on the reward prediction error ($\delta_{t,1}$), i.e. the difference between the value associated to the alternative and the obtained reward (Rescorla & Wagner (1972)), and a learning rate (α). Recall that stage 1 contains the alternatives, while stage 2 contains two possible states corresponding to each bus but no choices. In the first stage, the outcome obtained by choosing an alternative is the value associated to the subsequent state s plus the reward r_t. Then, the reward prediction error of choosing the alternative i and being taken to a state s where the reward r_t is obtained, can be estimated as the difference between the outcome of the choice and the value associated to the alternative i in time t:

$$\delta_{t,1} = QMF_{t,s} + r_t - QMF_{t,i}, \tag{1}$$

where $QMF_{t,s}$ is the value associated to the state $s \in S_2$ in trial t. Since after stage 2 just follows the reward (no further stages), then the reward prediction error and $\delta_{t,2}$ associated to the states s of S_2 (the buses) is given by the difference between the obtained reward r_t and the value associated to the state (bus) s in trial t:

$$\delta_{t,2} = r_t - QMF_{t,s} \tag{2}$$

After choosing an alternative i, the subject updates its value. If she obtains something better than expected ($\delta_{t,1} > 0$), then the associated value grows, otherwise it decreases. Then, for the next trial $t+1$, the subject updates the value of the previously chosen alternative:

$$QMF_{t+1,i} = QMF_{t,i} + \alpha_{t,1}\delta_{t,1}, \tag{3}$$

where $\alpha_{t,1}$ is the *learning rate*. The learning rate is a free parameter that indicates at which degree the subject's learning process is being influenced by the reward prediction error. It can be also interpreted, as the probability associated to replace the value $QMF_{t,i}$ by the value of the reward (see eq. **??**). We considered different learning rates in function of the reward prediction error, since disappointing experiences ($\delta < 0$) and satisfying experiences ($\delta > 0$) may influence in different magnitudes the learning process (Perez & Dickinson (2020). Then, $\alpha_{t,1} = \alpha^+$ if $\delta_{t,1} > 0$ and α^- in other case ($\alpha^+, \alpha^- \in [0, 1]$). Note that $\alpha_{t,1} = 0$ implies that the subject does not update the value associated to i, while $\alpha_{t,1} = 1$ implies that the subject completely replaces the previous value by the reward and the value of the state s ($QMF_{t+1,i} = QMF_{t,s} + r_t$). We estimated both α^+ and α^-.

Then, the subject updates the value of the bus corresponding to state s:

$$QMF_{t+1,s} = QMF_{t,s} + \alpha_{t,2}\delta_{t,2} \tag{4}$$

Similarly, the *learning rate* $\alpha_{t,1} = \alpha^+$ if $\delta_{t,2} > 0$ and α^- in other case. Note that $\alpha_{t,2} = 0$. If $\alpha_{t,2} = 1$, then $QMF_{t+1,s} = r_t$.

In this work we considered that the values associated to the non-chosen alternative and the non-visited state in stage 2 are updated with the same process but with a null reward. This is equivalent to define $QMF_{t+1,i} = QMF_{t,i}(1 - \alpha)$ if $y_{t,i} = 0$.

QMB estimation (Model Based algorithm)

On the other side, since each task comprises 200 trials, the MB component is calculated assuming that participants learned the transitions probabilities, which is a critical assumption, since represents the goal-directed nature of MB system. Recall that $QMB_{t,i}$ can be interpreted as an expected value, and for general purposes, as an expected utility. When assuming that subjects learned the transitions probabilities, we are implicitly assuming that they tried to understand the dynamics of the stochastic environment (i.e., the context of the task). $QMB_{t,i}$ can be formulated as depicted in eq. 5.

$$QMB_{t,i} = P(s_{blue}|i)QMB_{t,s_{blue}} + P(s_{red}|i)QMB_{t,s_{red}} \tag{5}$$

$P(s|i)$ is the probability of going to state (bus) s in stage S_2 after choosing the neighbor i. The values $QMB_{t,s}$ are the values that MB system associates to each state s of stage S_2. Similarly, as above, $QMB_{t,s}$ are learned using a *reward prediction error* (δ_t). In this case, the reward prediction error is the difference between the obtained reward and the value associated to the state (bus) that corresponds to time t. Recall that the reward prediction error indicates if the result of reaching bus s is better ($\delta_t > 0$) or worse ($\delta_t < 0$) than expected:

$$\delta_t = r_t - QMB_{t,s} \tag{6}$$

And then, on the next trial, the QMB value will be updated, generating $QMB_{t+1,s}$. Here, the learning rate (α_t) controls how much does the reward prediction error weights on the learning process, as shown in eq. 7.

$$QMB_{t+1,s} = QMB_{t,s} + \alpha_t\delta_t \tag{7}$$

In eq. 7 the learning rate also takes different values in function of the reward prediction error (δ_t), as explained before ($\alpha_t = \alpha^+$ if $\delta_t > 0$ and α^- in other case). Note that $\alpha_t = 0$ implies that $QMB_{t+1,s} = QMB_{t,s}$ (null learning), and $\alpha_t = 1$ implies that $QMB_{t+1,s} = r_t$ (the reward value is completely transferred to the value associated to the state s).

$QNET$ and model estimation

Finally, $QNET_{t,j}$ represents the mixture of both cognitive systems' values $QMB_{t,j}$ and $QMF_{t,j}$. As shown in eq. 8, $QNET_{t,j}$ is a convex combination of both values, where $\omega \in [0,1]$[3] is the *rate of MB behavior*.

$$QNET_{t,j} = (1-\omega)QMF_{t,j} + \omega QMB_{t,j} \tag{8}$$

If $\omega = 1$, the subject is fully goal-directed: the transition probabilities between states were learned and considered into decision-making. Therefore, if the subject is taken to a highly rewarding state, but because of the stochastic nature of the environment, does not get a reward, the probability of repeating the choice should not decrease, since it is understood that it was a random event with low probability of occurring. On the other hand, if $\omega = 0$, it means that subject's choices were driven solely by the rewards obtained at the final stage. In this case, and contrary to the previous case, if the subject is taken to a highly rewarding state, but does not get a reward, the probability of repeating the choice should decrease.

[3]To constrain ω, we estimated ω^*, with $\omega = exp(\omega^*)/(1+exp(\omega^*))$

Now, the probability of choosing the alternative i in the next trial $t+1$, is given by a softmax function (Sutton & Barto, 2020). We estimate the probability of choosing, in $t+1$, the same alternative chosen in t. Here underlies the assumption that participants decide binarily if stay with the same alternative or not, rather than choosing between two different alternatives. The probability is then formulated as follows:

$$P(y_{t+1,i} = 1|QNET_{t+1,i}, \beta) = \begin{cases} \frac{exp(\beta QNET_{t+1,i})}{exp(\beta QNET_{t+1,i})+1}, & y_{t,i} = 1 \\ \\ \frac{1}{exp(\beta QNET_{t+1,i})+1}, & y_{t,i} = 0 \end{cases} \tag{9}$$

In the above formulation, the parameter β is the *inverse temperature* (as it is called in psychology and neuroscience) (Sutton & Barto, 2020) or the *scale parameter* of RUM models. In the RUM interpretation of the term, the probability depends on QNET plus an error term. When that error has small variance, the scale, which is inversely proportional to the variance, is large, and the probability is almost deterministically determined by $QNET$. When the error has a large variance, the scale is small and the probability becomes close to 0.5, independent of $QNET$. Then, the joint probability of the observed decisions of a specific subject can be written as:

$$LogLik = \sum_{t=1}^{T} \sum_{i \in \mathcal{C}} log(P(y_{t,i} = 1)^{(y_{t,i}=1)}) \tag{10}$$

Finally, the parameters β, ω, α can be estimated by maximizing $LogLik$. For the estimation, all the Q-values were initialized at zero.

Exploration

As it was explained before, in the second part of the experiment, participants were offered a third option (another neighbor C), and were notified that a new bus had become available (a green bus). We are now interested in the rate of exploration of the new alternative. Drawing on previous studies (Daw et al., 2006), we consider a decision that choice of C in trial t was exploratory if and only if $QNET_{t,C} < QNET_{t,j}\ \forall j \neq C$. Otherwise, if the subject chooses the new alternative (C), but its $QNET$ is higher than the value of both A and B, we consider that as an exploitative decision. In other words, an exploratory choice occurs only if there is another, previously known and better valued alternative. Let $\tilde{y}_{t,i}$ indicate that the chosen alternative i had the minimum $QNET$, as depicted in eq. 11.

$$\tilde{y}_{t,i} = \begin{cases} 1 & QNET_{t,i} < QNET_{t,j} \forall j \neq i \wedge y_{t,i} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

We then defined the exploration rate ϕ_E of an alternative i for a specific subject as follows:

$$\phi_E(i) = \frac{\sum_{t=1}^{T} \tilde{y}_{t,i}}{T} \tag{12}$$

Which yields the proportion of exploratory choices during the second stage of the experiment. In this case, we are interested in the exploration of C, i.e. $\phi_E(C) = \frac{\sum_{t=1}^{T} \tilde{y}_{t,C}}{T}$.

The question at issue was the extent to which the parameter w is correlated with the rate of exploration in this second phase. We hypothesized that a higher ω would be associated with more exploration (or viceversa, that a lower ω would be associated with less exploration), reflecting that MF or habit-like strategies hinder exploratory behavior. The answer to this question is novel in both psychology and transportation.

4. RESULTS

Statistical analyses and computational modeling were performed using the R programming language (R Core Team, 2022) in a Microsoft (c) collab space. In the first phase of the experiment, our main objective was to estimate the parameter ω for each subject during the first stage of the task. Then, in the second phase, for each individual we calculated the exploration metric $\phi_E(C)$ when the new alternative was added to the set of options, and studied its correlation with the parameter ω. All parameters were estimated using maximum-likelihood through the maxLik function in R Henningsen & Toomet (2011); Daw et al. (2011). For some participants, the parameters were not identifiable because of invariant behavior and, for others, the parameters were identifiable, but the likelihood ratio test was non-significant (N=5). This left 29 participants for the final analysis.

Participants clearly showed an ordered preference (C $\succ$ B $\succ$ A) in the second phase of the experiment. More over, participants who preferred alternative C, obtained a better overall performance in the second part of the experiment (arrived on time more frequently).

As hypothesized, the exploration rate showed a significant correlation with the degree of MB behavior ($\rho = 0.493$, $p = 0.008$). We then separated participants by the median. Participants with ω above the median were defined as MB. Participants with ω below the median were defined as MF. 2B, shows the proportion of exploratory choices of alternative C for MB and MF participants when separating them using a median split. The exploration rate was on average 6.05% higher ($p < 0.001$) for MB participants than for MF. Also, when we modeled the learning process of probabilities in the MB system—rather than assuming that participants knew them from the outset of the experiment— we found that the correlation was also significant ($\rho = 0.652$, $p < 0.001$).

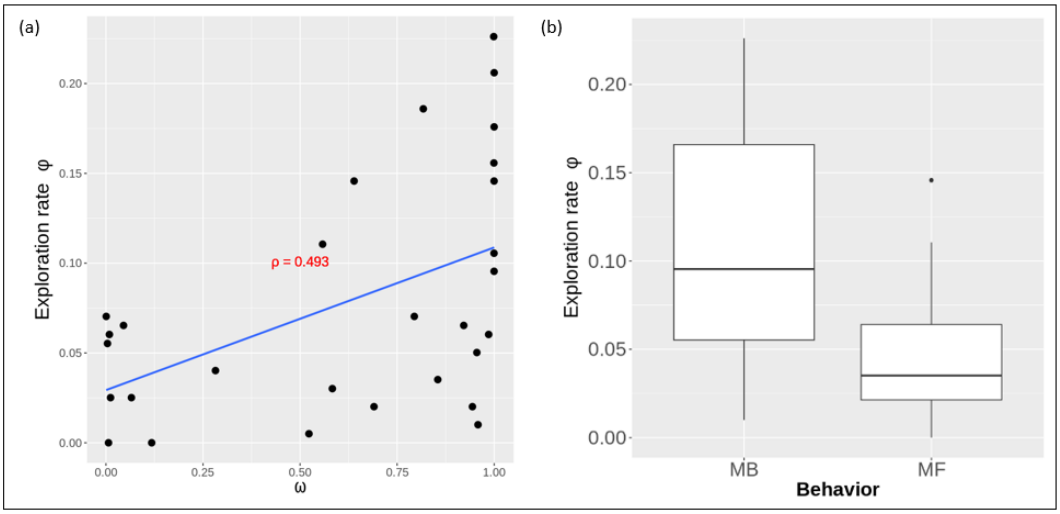

Figure 2: **Habits hinder exploration of new alternatives (a)** Exploration (ϕ) and MB behavior (ω) correlation. **(b)** Exploration rate distribution for MB and MF subjects. Participants were divided by a median split on ω.

5. DISCUSSION

Drawing on the literature on learning and decision making in psychology and neuroscience, we have obtained evidence suggesting that human exploration can be sub-optimal when participants rely on MF strategies. When participants are presented with a novel and better alternative in a markovian two-step task, their level of exploration was inversely related to their level of MF control.

While the concept of habit underlying MF strategies is not new in psychology and neuroscience, the exploration of this concepts in a transportation context is novel. The differentiation of behavioral strategies as MB and MF, offers a new perspective to conceptualize commuting habits. Rather than incorporating an inertia, model-free behavior allows for the modeling of the non-compensatory and automatic nature of habitual choices.

From a modeling perspective, inertia has shown to be successful in explaining commuter behavior (Valeri & Cherchi, 2016; Gao et al., 2020; Cantillo et al., 2007). However, modeling a utility-maximizing choice in a habitual behavior context is self-contradictory. When a subject behaves habitually, she just chooses automatically what is remembered to be of higher value. This phenomenon, despite the possibility of being modeled in a plethora of different ways, has been studied by computational neuroscientists for more than a decade. In this study, we draw on that literature to introduce the concepts to the travel behavior literature and shed light on why individuals showing habitual strategies may not explore new (and potentially better) alternatives. Until now, transportation researchers have addressed this problem by assuming that subjects ascribe a higher inertia to the old alternatives (Cantillo et al., 2007; Valeri & Cherchi, 2016; Gao et al., 2020). In this paper we offer an alternative explanation: habitual subjects are under MF control, which means they do not try to understand their environment to maximize utility, but just choose based on what they

remember to be a good choice.

In a real-life situation, a new transportation mode or a new available route may be underutilized just because travelers are behaving automatically and, therefore, are not reacting to the shock of a new (potentially better) travel alternative. Alongside transportation investments, it should be considered sufficiently strong behavioral measures to get people off the autopilot mode and incentivize the exploration of the new alternatives (Larcom et al., 2017). Otherwise, the investment would not have the desired welfare effects.

The reason why a MB subject would be more likely to explore than a MF or habitual subject are not clear. From a psychological standpoint, the question at hand is what factors would make an MB subject more inclined to search for potentially valuable information when uncertainty is high. One possibility is that MB subjects value information more than MF subjects, which would imply that the motivation to seek information is higher in MB subjects. Another, not completely unrelated possibility is that MB subjects tend to pay attention to novel events or stimuli, which is consistent with the idea that goal-directed behavior involves monitoring potential outcomes and options in the environment.

Our main hypothesis revolves around the potential inhibition of exploratory behavior due to the utilization of habitual MF strategies. However, the question that arises is the underlying psychological and computational mechanisms behind this phenomenon. Two perspectives have emerged in this regard. The first proposes that exploratory behavior is random, serving as an efficient strategy to thoroughly explore all available options in the environment. However, it seems improbable that this strategy was relevant to our subjects, given that the task was specifically designed to introduce a new option, from which information should be extracted in a goal-directed manner, with the aim of reducing uncertainty about its expected value. This aligns with the principles of directed exploration algorithms in computer science, wherein an agent's exploratory behavior is enhanced through the inclusion of uncertainty, novelty, or curiosity-based incentives (Gershman, 2018).

Outside of the lab, the sensitivity of people to changes in the environment (a proxy of habitual behavior) can be estimated by analyzing natural experiments like shocks in the transportation system (e.g. new metro lines), as has been studied elsewhere (Blase (1979); Arriagada et al. (2023); Ingvardson et al. (2023)). Transportation systems offer a natural laboratory to study repeated choices, with nowadays extensive data available thanks to intelligent transportation systems and automatic fare collection systems. This gives researchers in economics, psychology and transportation behavior the opportunity to probe behavioral patterns and factors common to different subjects which may make them more likely to behave in a habitual/goal-directed manner.

The present experimental framework could be used to address other relevant questions in transportation research. For simplicity, we have designed the reward function to yield only binary rewards (1 in the case of "arriving on time" and 0 in the case of "arriving late", similar to Fujii & Kitamura (2003)). This means that the value associated to each alternative is univariate, and that arriving on time or late has the same value for every subject. However, it is not difficult to conceive of an experiment where other attributes are presented in the task, which can be valued and encoded by subjects in a utility function (i.e. a continuous and subjective reward). In this way, as experimenters, we could potentially estimate the importance of attributes for each sub-

ject according to their own utility functions and how they are translated into choices, interacting with a learning rate and the MB/MF rate. This could have relevant impact in the welfare analysis. Moreover, by using RL algorithms it would be possible to model the learning of utilities (rather than Q-values) associated to different alternatives. In this regard, a latent classes model should permit the identification of a MF and a latent-based class. In such a model, the habitual behavior should be represented by a model that just considers last experiences to update utilities, while the goal-directed uses exogenous information (e.g. looking for information in traffic applications) to understand the environment and maximize utility. This also would allow the researcher to understand the nature of each class, by defining the probability of pertain to each class, for example, as a function of the socio-demographic characteristics or latent psychological variables as stress. However, the definition must be modified if another modeling framework is adopted. The definition here used is widely accepted and can be directly calculated in function of the $QNET$ (the mixture of the QMB and QMF). In a latent classes model, behavior mixture is done at the probability level and therefore could not be considered as a "net utility".

Which is the best way to model this dual behavior paradigm is an open question. Future research should tackle model comparisons and other techniques to differentiate habitual from goal-directed choices. Moreover, other cognitive and psychophysiological measures could provide further insight into the underlying neurological and psychological processes involved in comparing attributes during economic decision making (e.g. eye-tracking, physiological data) (see Hancock & Choudhury (2023)). Furthermore, psychological theories explored in travel behavior, e.g. the TPB, could interact with a dual system model, incorporating social norms and other people's attitudes.

The results of this study are limited by the sample characteristics. The sample size was limited by time constraints as the experiment was conducted in a physical laboratory in 12 one-hour blocks. The online replication of this study is currently in progress. This will allow a larger sample size with higher heterogeneity.

In sum, the main contributions of this article are: (1) the introduction of the concepts of MB and MF behavior to the transportation literature, which allows a better understanding of habits in commuting; (2) providing a modeling framework to estimate the degree of MF and MB for each individual; and more importantly (3) to show that MF behavior is negatively correlated with exploration, suggesting that habits may hinder the exploration of novel and potentially better alternatives. Even when more convenient alternatives might exist, habitual subjects might prefer to stay in *auto-pilot* mode, sticking with what they already know to be satisfying. Habits not only die hard, but they also make it hard to imagine better alternative worlds.

ACKNOWLEDGMENTS

This work was part of Bastián Henriquez-Jara's doctoral dissertation. This research was partially funded by an ANID-FONDECYT 1231584, ANID-PIA/PUENTE AFB220003 and ANID-FONDEF IT21I0059. Omar D. Perez is supported by ANID-SIA 85220023 and ANID-FONDECYT 1231027. We thank Weilun Ding for his valuable comments on the experimental design of this study.

REFERENCES

Arriagada, J., Guevara, A., Marcela, M., & Gao, S. (2023). An experiential learning-based transit route choice model using large-scale smart card data [unpublished manuscript].

Blase, J. H. (1979). Hysteresis and Catastrophe Theory: Empirical Identification in Transportation Modelling. **Environment and Planning A: Economy and Space**, 11 (6), 675–688.

Cantillo, V., De Dios Ortúzar, J., & Williams, H. C. (2007). Modeling discrete choices in the presence of inertia and serial correlation. **Transportation Science**, 41 (2), 195–205.

Daw, N. D., & O'Doherty, J. P. (2013). Multiple Systems for Value Learning. **Neuroeconomics: Decision Making and the Brain: Second Edition**, 393–410.

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., et al. (2011). Trial-by-trial data analysis using computational models. **Decision making, affect, and learning: Attention and performance XXIII**, 23 (1).

Dickinson, A., & Balleine, B. (1994, mar). Motivational control of goal-directed action. In N. J. Mackintosh (Ed.), **Animal learning & behavior** (Vol. 22, pp. 1–18). London: Academic Press. Retrieved from `http://doi.apa.org/psycinfo/1994-98574-002http://www.springerlink.com/index/10.3758/BF03199951` doi: 10.3758/BF03199951

Fujii, S., & Kitamura, R. (2003). What does a one-month free bus ticket do to habitual drivers ? An experimental analysis of habit and attitude change. , 81–95.

Gao, K., Yang, Y., Sun, L., & Qu, X. (2020). Revealing psychological inertia in mode shift behavior and its quantitative influences on commuting trips. **Transportation research part F: traffic psychology and behaviour**, 71, 272–287.

Gärling, T., & Axhausen, K. A. Y. W. (2003). Introduction : Habitual travel choice. , 1–11.

Gershman, S. J. (2015). Reinforcement learning and causal models. In W. M. (Ed.), **The oxford handbook of causal reasoning** (pp. 1–18). Oxford Library of Psychology.

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. **Cognition**, 173, 34–42.

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

Hancock, T. O., & Choudhury, C. F. (2023). Utilising physiological data for augmenting travel choice models: methodological frameworks and directions of future research. **Transport Reviews**.

Heckman, J. J. (1981). Statistical Models for Discrete Panel Data. In Charles F. Manski and Daniel L. McFadden (Ed.), **Structural analysis of discrete data and econometric applications** (pp. 113–178). The MIT Press.

Henningsen, A., & Toomet, O. (2011). maxlik: A package for maximum likelihood estimation in R. **Computational Statistics**, 26 (3), 443-458.

Ingvardson, J. B., Raveau, S., & Soza-Parra, J. (2023). Habit and shock effects in public transport: The case of metro line 6 in santiago using smart card data [unpublished manuscript].

Larcom, S., Rauch, F., & Willems, T. (2017). The benefits of forced experimentation: Striking evidence from the London underground network. **Quarterly Journal of Economics**, 132 (4), 2019–2055.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In Frontiers in Econometrics Academic Press (Ed.), **Zarembka** (pp. 105–1042).

Miller, E. J. (2020). **Travel demand models , the next generation : boldly going where no-one has gone before**. Elsevier.

Nussenbaum, K., Scheuplein, M., Phaneuf, C. V., Evans, M. D., & Hartley, C. A. (2020). Moving developmental research online: Comparing in-lab and web-based studies of model-based reinforcement learning. **Collabra: Psychology**, 6 (1), 1–18.

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. **Journal of neuroscience methods**, 162 (1), 8–13.

Perez, O. D., & Dickinson, A. (2020). A theory of actions and habits: The interaction of rate correlation and contiguity systems in free-operant behavior. advance online publication. **Psychological Review**.

R Core Team. (2022). R: A Language and Environment for Statistical Computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.r-project.org/`

Rasouli, S., & Timmermans, H. (2014). Applications of theories and models of choice and decision-making under conditions of uncertainty in travel behavior research. **Travel Behaviour and Society**, 1 (3), 79–90.

Rescorla, R., & Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. **Classical conditioning II: Current research and theory**, 2, 64–99.

Sutton, R., & Barto, A. (2020). **Reinforcement Learning** (Second ed.). The MIT Press Cambridge, Massachusetts.

Vaa, T. (2014). From Gibson and Crooks to Damasio: The role of psychology in the development of driver behaviour models. **Transportation Research Part F: Traffic Psychology and Behaviour**, 25 (PART B), 112–119.

Valeri, E., & Cherchi, E. (2016). Does habitual behavior affect the choice of alternative fuel vehicles? **International Journal of Sustainable Transportation**, 10 (9), 825-835.

Watkins, C., & Dayan, P. (1992). Q-learning. **Machine learning**, 8 (3-4), 279–292.

APPENDIX

Table A1: Glossary of main concepts of Reinforcement Learning

Concept	Definition
Value	Mental representation of an alternative or action. It allows for the comparison and oredering of alternatives. It can be understood as a utility, but which is constantly learned rather than computed as a function of the attributes of the alternatives.
Reward	Outcome of an action. It is commonly discretly codded (1 for positive reward, -1 for negative reward and 0 for a non-rewarded action).
Stage	Components of the structure of a learning process. Each stage may comprise a set of possible actions or alternatives, and each of them make take to different stages.
Transition	In a markovian decission process, the transitions are the connection between stages.
Q-learning	A particular RL alogithm, developed by Watkins & Dayan (1992), in which the values are updated according to the RPE and a learning rate.
Reward prediction error (RPE)	Is is the difference between the value associated to an alternative and the obtained reward after choosing it.
Learning rate	The rate at which the RPE is incorpored in to the learning process of the value associated to an alternative.
Model-free	Cognitive system that represents the learning process of a subject with habitual or automatic behavior. A subject under habitual behavior, is said to be "controlled by" model-free system.
Model-based	Cognitive system that represents the learning process of a subject with goal-directed (utility maximizer) behavior. A subject under goal-directed behavior, is said to be "controlled by" model-based system.