American Psychological Association

Journal of Experimental Psychology: General

0096-3445/19/$12.00

2019, Vol. 148, No. 6, 977–993
http://dx.doi.org/10.1037/xge0000546

Should We Control? The Interplay Between Cognitive Control and Information Integration in the Resolution of the Exploration-Exploitation Dilemma

Irene Cogliati Dezza and Axel Cleeremans
Université Libre de Bruxelles

William Alexander
Ghent University and Florida Atlantic University

In their daily decisions, humans and animals are often confronted with the conflicting choice of opting either for a rewarding familiar option (i.e., exploitation) or for a novel, uncertain option that may, however, yield a better reward in the near future (i.e., exploration). Despite extensive research, the cognitive mechanisms that subtend the manner in which humans solve this exploration-exploitation dilemma are still poorly understood. In this study, we challenge the popular assumption that exploitation is a global default strategy that must be suppressed by means of cognitive control mechanisms so as to enable exploratory strategies. To do so, we asked participants to engage in a challenging working memory task while performing repeated choices in a gambling task. Results showed that manipulating cognitive control resources exclusively hindered participants' ability to explore the environment in a directed, intentional manner. Moreover, under certain scenarios, adopting exploitative strategies was also dependent on the availability of cognitive control resources. Additional analyses using a recent computational model of information integration suggests that increasing cognitive load specifically interferes with the ability to combine reward and information in order to inform choices. Our results shed light on the cognitive mechanisms that underpin the resolution of the dilemma and provide a formal foundation through which to explore pathologies of goal-directed behavior.

Keywords: exploration-exploitation dilemma, informative value, reinforcement learning, cognitive control, adaptive behaviors

Understanding the exploration-exploitation dilemma is widely taken to be one of the main challenges in the domain of adaptive control and behavior (Cohen, McClure, & Yu, 2007). The dilemma refers to the fact that when facing a choice, one may either choose to stick with what we know (familiar rewarding outcomes) or engage in the risky exploration of unknown regions of the decision space. To better picture this phenomenon, imagine that it is a nice day in your city. You are walking around downtown in search of a pleasant place to eat. A good strategy would be to choose your favorite restaurant, because the likelihood that you will find it satisfying is very high. However, new dining rooms have recently opened in town. Do you select the restaurant that you know you will enjoy, or do you select another restaurant that you have never tried before, potentially finding either a new favorite or profound disappointment? Thus, the exploration-exploitation trade-off is a dilemma precisely because it involves addressing a challenging conflict between maximizing reward and maximizing information. Solving it is necessary in order to flexibly adapt to environments that are often both uncertain and dynamic. Because all cognitive agents have to somehow address this challenge, the exploitation-exploration dilemma is ubiquitous and has relevance for many organisms and for many types of decisions.

Although extensive research on the exploration-exploitation dilemma has been conducted over the last decades in different scientific domains (e.g., artificial intelligence, animal foraging, and neuroscience), a complete understanding of the underlying mechanisms involved in the resolution of the dilemma is still lacking. In the most popular framework (Cohen et al., 2007; Daw, O'Doherty, Dayan, Seymour, & Dolan, 2006), the dilemma is considered as a dual-process in which exploitation is a default

This article was published Online First January 21, 2019.

Irene Cogliati Dezza and Axel Cleeremans, Centre for Research in Cognition & Neurosciences (CRCN), ULB Neuroscience Institute (UNI), Université Libre de Bruxelles; William Alexander, Department of Experimental Psychology, Ghent University, and Center for Complex Systems and Brain Sciences, Florida Atlantic University.

Irene Cogliati Dezza is a researcher supported by Fonds de la Recherche Scientifique (F.R.S.-FNRS) grant (Belgium). Axel Cleeremans is a research director with the F.R.S.-FNRS (Belgium). William Alexander was supported in part by FWO-Flanders Odysseus II Award G.OC44.13N. Irene Cogliati Dezza designed and carried out the experiment. Irene Cogliati Dezza performed the analysis of the data and the model analysis. Irene Cogliati Dezza and William Alexander discussed and interpreted the data. Irene Cogliati Dezza, Axel Cleeremans, and William Alexander wrote the manuscript. The results reported in this article were presented at the Belgian Association for Psychological Science (BAPS) annual meeting in 2018.

Correspondence concerning this article should be addressed to Irene Cogliati Dezza, Centre for Research in Cognition and Neurosciences (CRCN), ULB Neuroscience Institute (UNI), Université Libre de Bruxelles, Avenue F.D. Roosevelt, 50, 1050 Bruxelles, Belgium. E-mail: icogliat@ulb.ac.be

strategy, and it appears to dominate choice behavior because of its association with stronger reward histories. Following this framework, modifying behavior in an adaptive manner through exploration thus requires overriding the exploitative strategies that tend to dominate the decision process by its stronger association with rewards. To overcome this dominance, behavioral/cognitive control processes might play a central role (i.e., inhibition) in enabling the switch to exploratory strategies (Cohen et al., 2007; Daw et al., 2006). Cognitive control is the ability to coordinate sensory information and actions so as to align them to internal states or intentions (Koechlin, Ody, & Kouneiher, 2003), and is required when the mapping between sensory inputs and actions is rapidly changing or weakly established relative to other existing stimulus–response associations (Miller & Cohen, 2001). Top-down control mechanisms could therefore be the core process that underpins exploratory behavior by enabling the continuous monitoring of the need for behavioral adjustments and by implementing new goal-directed behaviors (Cohen et al., 2007; Daw et al., 2006). The "behavioral control" framework was introduced to explain activity in prefrontal regions (i.e., frontopolar cortex), known to be involved in cognitive control (Mars, Sallet, Rushworth, & Yeung, 2011) during exploratory decisions (Daw et al., 2006). Subsequent evidence has confirmed the core involvement of higher cognitive-control functions in exploration (Badre, Doll, Long, & Frank, 2012; Cavanagh, Figueroa, Cohen, & Frank, 2012; Frank, Doll, Oas-Terpstra, & Moreno, 2009).

To understand precisely how cognitive control is related to choice behavior in the exploration-exploitation dilemma, it is important to note that, under the framework outlined in the previous paragraph, *exploitation* specifically refers to choosing the option that maximizes a (reward) prediction. Exploration, on the other hand, is an umbrella term that encompasses different type of strategies, essentially *random* and *directed* exploration (Wilson, Geana, White, Ludvig, & Cohen, 2014). The concept of *random exploration* derives from reinforcement learning (RL) theory (Sutton & Barto, 1998), wherein exploration is merely the product of noise in the response-generation process. Under this scenario, a decision maker who learns to maximize a numerical reward signal may nevertheless make choices associated with lower reward values (exploration) because of a noisy response. In contrast, the concept of directed exploration derives from optimal decision-making theories, which take exploration to be an explicit, goal-directed strategy (Gittins & Jones, 1974). In *directed exploration*, an animal "directs" exploration toward uncertain options, thus increasing its understanding of the surrounding environment through gaining new information. Thus, the absence of information is the main driving factor in this subtype of exploration behavior.

Whether humans use information to direct their exploratory behaviors has been a matter of intense discussion over the last decade, and a number of findings have suggested that this is not the case (Daw et al., 2006; Payzan-LeNestour & Bossaerts, 2011). However, this view has been challenged recently in studies using alternative paradigms that controlled for the availability of information in the environment, suggesting that humans may adopt both random and directed exploration (Wilson et al., 2014), the hidden mechanisms of which relate to the integration of reward and information into choice values (Cogliati Dezza, Yu, Cleeremans, & Alexander, 2017). Although based on a common exploratory drive, the two exploratory strategies showed different neural substrates (Warren et al., 2017; Zajkowski, Kossut, & Wilson, 2017), different age-related development (Somerville et al., 2017), and they react differently to changes in reward contingencies (Cogliati Dezza et al., 2017). Thus, the dilemma does not seem to be a unitary binary process but instead a class of problems spanning different scales (Cohen et al., 2007). Following this recent perspective, the dilemma is represented as a continuum (Mehlhorn et al., 2015) on which many behaviors fall in the extremes (e.g., choosing the highest valuable option or the most uncertainty option), whereas others might fall somewhere in between (choosing a moderately valuable option associated with some uncertainty). Behavior at these intermediate points on the continuum is less amenable to interpretation, and controlled behavioral paradigms are required (Wilson et al., 2014). Different cognitive mechanisms may therefore underlie the resolution of the dilemma, and the ability of a decision maker to deploy different exploratory strategies may depend on the availability of sufficient cognitive control resources (Otto, Knox, Markman, & Love, 2014). However, a new framework that attempts to integrate these new advances in understanding the exploration-exploitation dilemma and its underlying cognitive mechanisms is still lacking.

Motivated by the behavioral control hypothesis of exploratory behavior and by recent understanding over the resolution of the exploration-exploitation dilemma in humans, we consider whether cognitive control processes might modulate the resolution of the exploration-exploitation dilemma using a mixture of exploratory strategies (i.e., random and directed exploration). We investigated this hypothesis using a variant of bandit tasks that has previously been used to disentangle both random and directed exploratory strategies (Cogliati Dezza et al., 2017; Wilson et al., 2014). Bandit tasks are a family of RL problems in which, for each trial, participants must choose among a set of slot machines (or "bandits") with the goal of maximizing the total reward over a sequence of trials (Robbins, 1952). This new version of the bandit task used a two-phase gambling task in which, for each game, participants were initially instructed as to which options to choose (*forced-choice task*), after which they were free to choose between options (*free-choice task*) so as to maximize their final gain. By adding a forced-choice task on the top of the standard bandit task, the information participants had about the payoffs of each option was controlled, thereby enabling the identification of the two exploratory strategies in the first free-choice trial of each game (Wilson et al., 2014). In the current study, we additionally manipulated cognitive control resources by asking participants to engage in a challenging working memory task (Konstantinou & Lavie, 2013) while performing the sequential decision-making task. Under the behavioral control hypothesis, depletion of cognitive control resources should lead to a more pronounced expression of processes that operate independently of control, such as exploitation, whereas behaviors that require control—such as exploration—should be attenuated. In order to investigate the effect of cognitive load manipulation on the learning and decision-making components of the dilemma, we developed a computational model that is capable of capturing participants' behavior on the new version of the bandit task by associating a value with information on top of the standard reward-based RL formulation (Cogliati Dezza et al., 2017). Applying a computational model in this context will help in

understanding the underlined mechanisms affected by cognitive control manipulation, which might be not accessible with a "pure" behavioral analysis.

Method

Participants

Twenty-five young adults participated in this study (20 women; aged 18–24 years, mean age = 19.6). Based on a previous study (Cogliati Dezza et al., 2017), a power analysis suggested a sample size of 24 and power of 0.999. Participants were students at the Faculty of Psychology (Université libre de Bruxelles) and received credits for their participation to the study. The entire group belonged to the Belgian French-speaking community. The experiment was approved by Faculty of Psychology Ethics Committee. Informed consent was obtained from all participants prior to the experiment.

Procedure

Bandit task. To investigate the effect of cognitive control on the exploration-exploitation dilemma, we asked participants to perform 128 independent games of a new version of the multi-armed bandit task (see Figure 1) that has already been shown to elicit both random and directed exploratory strategies (Cogliati Dezza et al., 2017; Wilson et al., 2014). As in standard bandit tasks, in this version, participants chose among options with the goal of maximizing the total reward over a sequence of trials. When selected, each option provides a reward (generated from a hidden distribution) that informs participants about the desirability of each alternative. Contrary to standard bandit tasks, for each game participants performed a forced-choice task followed by a

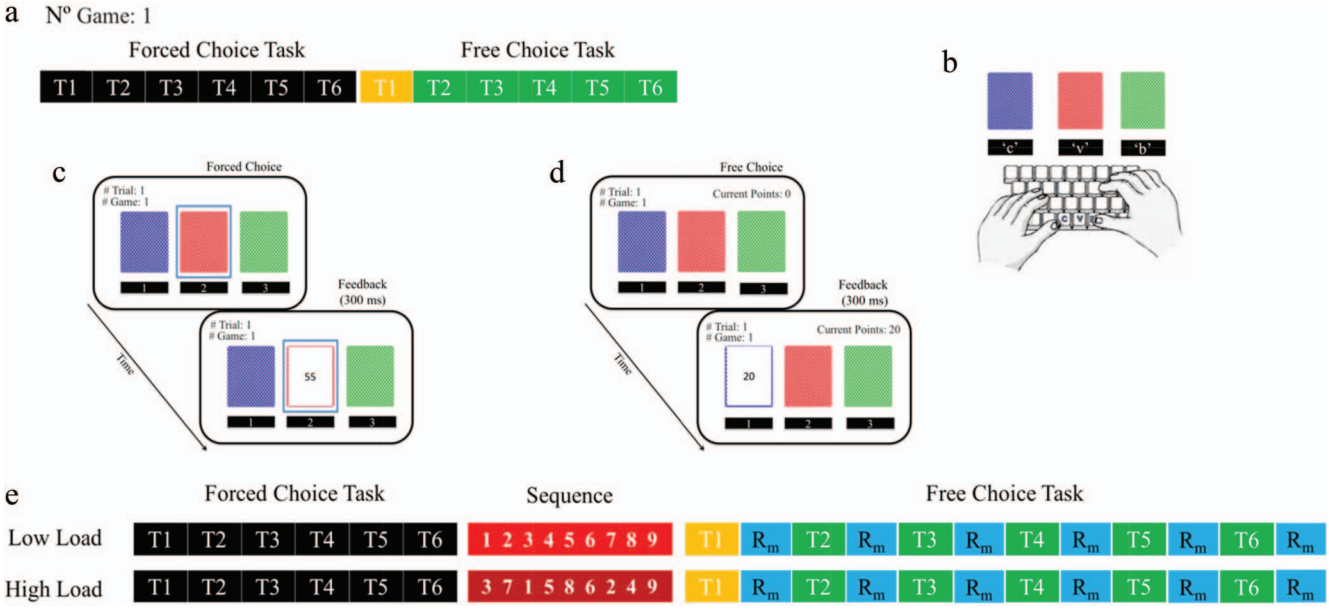

Figure 1. Behavioral paradigm. (a) Organization of games and trials. For each game, participants faced six consecutive trials of the forced-choice task and between one and six trials of the free-choice task. In the first free-choice trial (in yellow or T1), reward and information are orthogonalized, enabling the distinction between random and directed exploration. The number of free-choice trials was exponentially distributed such that a higher proportion of games allowed subjects to make six free choices. (b) Choices: Participants indicated their choices using the forefinger, middle finger, and ring finger and pressing the keyboard keys "C," "V" and "B," respectively. (c) Forced-choice task: Three decks of cards were displayed on the screen (a blue, a red, and a green deck) and participants were forced to choose a preselected deck (outlined in blue in the figure). After selecting the deck, the card turned and revealed the points associated with the selected option, between 1 and 100 points. At this stage, the points displayed to participants were not added to their total score. (d) Free-choice task: Participants made their own decisions among the same three decks of cards displayed during the forced-choice task. After each trial, the points displayed on the screen were added to the participants' total score and participants were instructed to attempt to maximize the total points earned at the end of the experiment. (e) Cognitive load manipulation: Before the first trial of the free-choice task, a sequence of nine digits was displayed on the screen. During the Low Load condition, the digits were presented in fixed numerical order (i.e., "123456789") for 500 ms. On the contrary, during the High Load condition, the digits were presented in random order (i.e., "371586249") for 2,000 ms, and a new sequence was generated for each game. After each free-choice trial, a digit (randomly selected from the nine-digit sequence) was displayed to participants who needed to report ("R_m": memory response) the number that followed the presented number in the previous nine-digit sequence presented before the first free-choice trial. T = trials. See the online article for the color version of this figure.

free-choice task (Wilson et al., 2014; Figure 1a). During the forced-choice task, participants were only allowed to select options that had been preselected by the computer (Figure 1c), whereas during the free-choice task, participants were able to make their own choices in view of maximizing their final score (i.e., the amount of points earned throughout the game; Figure 1d). Contrary to the first version of this paradigm (Wilson et al., 2014), information regarding the points earned following a choice did not remain visible following a feedback in order to allow learning to influence participants' choices (Cogliati Dezza et al., 2017; Zajkowski et al., 2017). Each game was composed of six consecutive forced-choice trials and from one to six free-choice trials (Figure 1a). The number of free-choice trials was manipulated so that participants were unable to predict the length of the free-choice task (Cogliati Dezza et al., 2017) and to adjust their choices accordingly (Wilson et al., 2014).

In this version, options were represented as decks of cards and were placed on the left (blue deck), right (green deck), and central (red deck) side of the computer screen (Figure 1b). The use of three options allowed us to discern between the strategic use of random and directed exploration (Cogliati Dezza et al., 2017) without manipulating the prior knowledge participants had about horizon (i.e., the total number of trials participants will experience in a game), as in previous versions (Krueger, 2017; Wilson et al., 2014). In particular, if choice probabilities for the two nonexploitative options are equal, then exploratory behavior is entirely driven by random exploration. On the contrary, if the choice probability is different from chance, then choices are partially driven by directed exploration. Participants indicated their choices using the buttons "C," "V," and "B" of the computer keyboard (Figure 1b). After each choice, the card was turned to reveal the points earned by the participant for selecting that deck. Participants could obtain between 1 and 100 points for each trial, and the number of points earned for selecting a deck was sampled from a truncated Gaussian distribution with a standard deviation of 8 points (the standard deviation was equal for the three decks). The generative mean of each deck was set to 30 and 50 points and adjusted by ± 0, 4, 12, and 20 points to avoid the possibility that participants might be able to distinguish the generative mean for a deck after a single observation (i.e., the generative means ranged from 10 to 70 points). As in our previous study (Cogliati Dezza et al., 2017), the three decks of cards had the same generative means in 50% of the games (equal reward condition) and different means in the rest of the games (unequal reward condition); the intent of the different reward conditions in our previous study was to examine the influence of reward context on exploration and exploitation. Although not the primary focus of this study, reward context effects reported in our previous study were also observed here ($p < 10^{-3}$) replicating our previous work. However, in the present study, the effect of reward context was not modulated by the cognitive control manipulation. For this reason, the results concerning reward context will be not discussed further. The means of the generative Gaussian function were stable within a game and varied between games. Participants were informed that the decks of cards did not change during the same game but were replaced by new decks at the beginning of each game. However, they were not informed of the reward manipulation and the underlying generative distribution we adopted.

As in previous versions of this paradigm, during the forced-choice task, we manipulated the information about the decks of cards acquired by participants (i.e., the number of times each deck of cards was played). For each game, participants were forced to either choose each deck 2 times (equal information condition) or to choose one deck 4 times, another 2 times, and never for the remaining deck (unequal information condition). The information manipulation guarantees the orthogonalization of reward and information, thus allowing the distinction of random and directed exploration in the first free-choice trial of each game (Wilson et al., 2014). In 50% of the games, participants played with the equal information condition. The order of card selection was randomized in both information conditions as well as the appearance of equal and unequal information condition.

Cognitive control manipulation. Cognitive control resources were manipulated by asking participants to carry out a concurrent working memory task during the free-choice task. Specifically, we adopted Konstantinou and Lavie's (2013) procedure, which has been shown to selectively interfere with cognitive control processes (Baddeley, Emslie, Kolodny, & Duncan, 1998; D'Esposito, Postle, Ballard, & Lease, 1999). Prior to the beginning of the free-choice task, a sequence of nine digits appeared on the screen (Figure 1e). Participants were asked to memorize and retain the sequence until the end of the game. After each free-choice trial, a single memory probe digit was presented at fixation until a response was given. The probe was equally likely to be any of the first eight digits of the memory set. The participants' task was to report the digit following the probe in the memory sample (e.g., if the memory set was "123456789" and the probe was "3," the correct response would be "4"). The probe was displayed on the screen and participants pressed the key corresponding to the selected digit.

In order to investigate the role of cognitive control resources on the exploration-exploitation dilemma, participants were exposed to two different conditions: High Load versus Low Load. In the High Load condition, the digits were presented in random order (e.g., "371586249") for 2,000 ms, and a new sequence was generated for each game. In the Low Load condition, the digits were presented in fixed numerical order (i.e., "123456789") for 500 ms. Participants performed the two conditions on two different days, with order randomized and counterbalanced (half of participants performed the High Load condition on the first day and the Low Load condition the second day, and vice versa). Performance on the memory task was adopted as an inclusion criterion for the statistical analysis (see Results section). Because of technical problems, two participants failed to complete the entire 128 games in either the High Load or the Low Load condition, but their data were included anyway because only a few games were lacking (one participant played 124 games of High Load condition and the other 123 of the Low Load condition), and removing those participants did not affect the main results.

Computational Models

To investigate the hidden mechanisms involved in the resolution of the exploration and exploitation dilemma under cognitive load, we adopted a previously implemented version of a RL model that learns reward values for each trial and incorporates a mechanism reflecting the knowledge gained about each deck during previous

experience—the gamma-knowledge reinforcement learning model (gkRL). The gkRL model is able to reproduce participants' behavior on the behavioral paradigm described in the previous paragraph (Cogliati Dezza et al., 2017). Specifically, compared with a standard RL model, this model is able to reproduce participants' directed exploratory strategies in scenarios in which options are not sampled at the same rate.

In each trial, a simple δ learning rule (Rescorla & Wagner, 1972) is used to compute the expected reward value $Q(c)$ for each deck of cards c (left, central, or right), using the following equation:

$$Q_{t+1,\,j}(c) = Q_{t,j}(c) + \alpha \times \delta_{t,j} \tag{1}$$

where $Q_{t,j}(c)$ is the expected reward value for trial t and game j; $\delta_{t,j} = R_{t,j}(c) - Q_{t,j}(c)$ is the *prediction error*, which quantifies the discrepancy between the predicted outcome and the actual outcome obtained at trial t and game j. The expected reward $Q_{t,j}(c)$ is updated using the (1) only if an outcome from the deck c is observed; otherwise, $Q_{t+1,\,j}(c) = Q_{t,\,j}(c)$. Considering that participants were told that each game was independent from the others, Q_0 is initialized at the beginning of each game (Khamassi, Enel, Dominey, & Procyk, 2013) and set to the global estimate of Q (~40 points; Cogliati Dezza et al., 2017).

Additionally, gkRL tracks information gained from each deck based on how often it is selected, as follows:

$$I_{t,\,j}(c) = \left(\sum_{1}^{t} i_{t,j}(c)\right)^{\gamma}$$

where

$$i_{t,j}(c) = \begin{cases} 0, & choice \neq c \\ 1, & choice = c \end{cases} \tag{2}$$

and where $I_{t,j}(c)$ is the amount of information associated with the deck c at trial t and game j, and $I_{t,j}(c)$ is computed by including an exponential term γ that defines the degree of nonlinearity in the amount of observations obtained from options after each observation. Gamma (γ) is constrained to be more than zero. Each time deck c is selected, $i_{t,j}(c)$ takes value of 1, and 0 otherwise. On each trial, the new value of $i_{t,j}(c)$ is summed to the previous $i_{t-1,1:j}(c)$ values, and the resulting value is raised to γ, resulting in $I_{t,j}(c)$. For example, after six trials of the forced-choice task, if one option has never been selected, $I_{t,j}(c)$ has value zero, whereas in the case that one option is selected 4 times, $I_{t,j}(c)$ has the value $4\hat{\gamma}$. The parameter γ adds nonlinearity to the information term (Cogliati Dezza et al., 2017), the intuition being that additional samples do not contribute equally to the amount of information a subject has about an option (e.g., sampling an option you have never observed is far more informative than sampling an option you have observed 100 times previously).

Before selecting the appropriate option, gkRL subtracts the information gained, $I_{t,j}(c)$, from the expected reward value, $Q_{t,j}(c)$:

$$V_{t,j}(c) = Q_{t,\,j}(c) - I_{t,\,j}(c) \times \omega \tag{3}$$

where $V_{t,j}(c)$ is the final value associated with deck c. Here, information accumulated during the past trials scales values $V_{t,j}(c)$, so that increasing the number of observations of one option decreases its final value. In other words, when one option is overselected, $I_{t,j}(c)$ becomes larger resulting in lower $V_{t,j}(c)$ On the contrary, if one option is never selected, $I_{t,j}(c)$ is zero and $V_{t,j}(c) = Q_{t+1,\,j}(c)$. In Equation 3, ω is the information weight and determines the degree by which the model integrates information into choice values. In order to generate choice probabilities based on expected reward values, the model uses a softmax choice function (Daw et al., 2006; Humphries, Khamassi, & Gurney, 2012; Wilson & Niv, 2012). The softmax rule is expressed as

$$P(c/V_{t,j}(c_i)) = \frac{\exp[\beta \times V_{t,j}(c)]}{\sum_i \exp[\exp\beta \times V_{t,j}(c_i)]} \tag{4}$$

where β is the inverse temperature that determines the degree to which choices are directed toward the highest rewarded option. With higher β, the model mainly selects options associated with higher choice value, whereas with lower β, the model's choices are more random.

The gkRL model can be informative concerning the effect of cognitive load on the dilemma in two ways. First, it can help to distinguish whether cognitive load effects on exploration are driven by information computation (ω and γ) or whether they are instead driven by changes in choice variability (β). Second, if changes are driven by alterations in information computation, the model can help to distinguish whether these are driven by changes in information integration (ω) or by changes in the way information availability decays with time (γ).

Model Fitting and Model Comparison

To estimate the model's parameters α, β, and ω, γ, we collected trial-by-trial participants' choices in both the High and Low Load conditions (Table 1, Table 2). During the fitting procedure, the objective function—the negative log likelihood, $\sum_{j=1}^{j=128} \log(P_j(c))$, for each participant under both load conditions was computed and then minimized using MATLAB and Statistics Toolbox (Release 2015b) function fminsearchbnd (which is exactly as fminsearch but does not search outside the fixed boundaries). The boundaries adopted were as follows: α]0,1[, β]0, 10], ω [−300, 300], γ]0, 12]. To increase the likelihood of finding a global rather than a local optimum, fminsearchbnd was iterated with 15 randomly chosen starting points. The fitting procedure was validated by running a recovery analysis: The gkRL model was simulated on the task using the retrieved parameter estimates to generate synthetic behavioral data, and then the fitting procedure was applied to the synthetic data in order to check whether previously estimated parameters were indeed recovered ($r^2 > 0.4$). Likewise, we checked the model comparison outcome by computing a confusion matrix and checking whether data generated from a model was indeed best explained by that model.

Statistical Analysis

Statistical analysis was performed using RStudio (https://www.rstudio.com/); the functions and packages adopted are reported in the Results section. To determine whether and how manipulating cognitive control affected participants' decision strategies, we conducted repeated-measures ANOVA analyses. When violations of parametric tests were indicated, nonparametric tests were performed; p values of less than 0.05 were considered significant.

Table 1
Model Fit Results: First Free-Choice Trials

Participant subject number	Low Load					High Load				
	α	β	ω	γ	Log (γ)	α	β	ω	γ	Log (γ)
1	.870	.142	21.754	.002	–6.458	.562	.145	12.740	.520	–.653
2	.464	.283	−10.186	.000	−19.612	.566	.094	1.017	.202	−1.601
3	.371	.337	1.979	.289	−1.240	.547	.080	2.667	1.313	.272
4	.587	.186	.000	10.386	2.340	.482	.297	−2.314	.359	−1.024
5	.000	6.392	−.067	.595	−.518	.000	7.500	.000	.919	−.084
6	.724	.132	9.246	.000	−19.089	.599	.132	−8.941	.110	−2.209
7	.254	.173	6.058	.000	−25.509	.109	.201	.000	7.882	2.065
8	.463	.131	18.008	.321	−1.137	.011	1.367	.230	.000	−21.745
9	.258	.042	−3.445	.000	−19.636	.455	.000	148.72	2.125	.754
10	.190	.400	.011	4.429	1.488	.004	4.693	−.048	.811	−.209
11	.455	.120	12.844	.000	−25.225	1.000	.040	−13.591	.558	−.583
12	.270	.283	6.120	.310	−1.170	.343	.072	2.610	.000	−21.931
13	.385	.077	2.828	.664	−.410	.661	.084	9.516	.251	−1.383
14	.502	.076	28.386	.000	−21.935	.345	.351	1.306	1.354	.303
15	.510	.248	9.855	.326	−1.119	.422	.165	10.952	.291	−1.233
16	.698	.096	−4.486	.000	−21.560	.563	.102	−26.370	.145	−1.931
17	.413	.134	−16.646	.000	−20.527	.432	.150	−21.224	.000	−22.509
18	.564	.114	14.187	.000	−24.509	.004	2.181	−.337	.803	−.219
19	.463	.219	−1.686	.506	−.681	.556	.142	−1.872	.742	−.298
20	.536	.462	14.166	.298	−1.211	.733	.187	.000	7.368	1.997
21	.143	.530	2.390	1.090	.086	.534	.140	5.736	.000	−20.851
22	.004	3.518	.000	11.000	2.398	1	.004	−201	.000	−25.761
23	.054	.771	1.145	.706	−.348	.569	.140	.000	9.404	2.241
24	1.000	.065	.000	10.000	2.303	.002	3.278	−.057	.976	−.024
25	.468	.215	8.107	.519	−.657	.897	.081	10.165	.000	−20.222
Total	.426 (.249)	.606 (1.38)	4.82 (9.89)	1.66 (1.40)	−8.16 (10.79)	.456 (.3)	.865 (1.8)	−2.81 (51.76)	1.44 (2.62)	−5.47 (9.68)

Note. Estimated parameters for each subject using the gamma-knowledge reinforcement learning model during High Load and Low Load conditions. Group averages of the estimated parameters are also reported. Group standard deviations are reported in parentheses.

Results

In this section, we first report the results concerning the cognitive load manipulation we adopted and its effects on participants' performance. Subsequently, we examine the interaction between cognitive load manipulation and decision strategies. Lastly, we investigate the possible hidden mechanisms affected by manipulating cognitive/behavioral control mechanisms.

Working Memory Task

First, we explored the effect of the cognitive load manipulation on memory accuracy. To do so, trial-by-trial correct memory responses were collected. A Wilcoxon signed-ranks test on the average value of subjects' overall correct memory responses revealed a significant difference between High Load ($M = 0.494$, $SD = 0.12$) and Low Load ($M = 0.986$, $SD = 0.012$; $p < 10^{-8}$, $r = .874$), indicating that, as expected, increasing memory load affected participants' performance on the working memory task (Figure 2a). Because it can be assumed that participants who scored at chance level on memory performance were not reliably engaged in the memory task, accuracy on the memory task was used as an inclusion criterion for further statistical analysis. A one-sample t test on correct memory responses revealed a significant difference between the High Load condition and chance level (12.5%), $t(24) = 15.29$, $p < 10^{-14}$, $d = 4.33$, suggesting that participants, on average, were actively engaged in the working memory task. Additionally, we investigated whether each participant performed at an above-chance level by applying a one-sample sign test on participants' correct memory responses in the High Load condition. Results revealed that each participant scored above chance level ($p < 10^{-6}$). Following this result, every participant was included in the subsequent analysis.

Cognitive Load Manipulation

To check whether the cognitive load manipulation affected cognitive control processes by increasing dual-task interference, we measured choice reaction times (RTs) during the free-choice task of both High and Low Load conditions (Figure 2b). A paired t test on RTs revealed slower RTs during the High Load condition ($M = 1005$ ms, $SD = 468$ ms) compared with the Low Load condition ($M = 483$ ms, $SD = 145$ ms), $t(24) = 6.19$, $p < 10^{-6}$, $d = 1.24$, suggesting that less cognitive control resources were available to participants during the High Load manipulation.

Performance

We also examined whether the cognitive load manipulation affected the way participants performed the gambling task. Here, *performance* refers to the ability to play strategically during the task in order to maximize the total gain. To do so, we computed the probability of choosing the deck with the highest average of points obtained in the previous trials (overall performance) during the entire free-choice task under all reward conditions in both High

Table 2
Model Fit Results: All Free-Choice Trials

Participant subject number	Low Load					High Load				
	α	β	ω	γ	Log (γ)	α	β	ω	γ	Log (γ)
1	.640	.119	19.633	.070	–2.659	.570	.114	15.735	.072	–2.631
2	.562	.112	−.465	1.403	.338	.339	.121	−.018	3.718	1.313
3	.475	.168	4.188	.000	−21.921	.590	.070	5.611	.000	−21.464
4	.446	.254	.958	2.943	1.079	.412	.238	−.550	1.952	.669
5	.000	2.543	−.040	.000	−19.081	.306	.010	−37.732	.674	−.395
6	.698	.124	4.107	.000	−21.476	.580	.072	−4.313	1.064	.062
7	.397	.103	9.887	.000	−22.202	.336	.049	−.075	3.504	1.254
8	.627	.110	26.525	.000	−27.211	.006	1.500	.041	12.000	2.485
9	.002	2.958	−.005	11.847	2.472	.096	.009	6.579	1.546	.435
10	.526	.180	−.001	4.443	1.491	.210	.090	−2.064	.636	−.452
11	.388	.123	−.033	3.179	1.157	.132	.131	−5.146	.682	−.383
12	.443	.225	9.320	.000	−22.414	.451	.067	−.072	3.044	1.113
13	.513	.067	−.005	4.412	1.484	.258	.125	4.366	.000	−21.627
14	.597	.091	17.966	.007	−4.918	.296	.234	−.031	2.916	1.070
15	.467	.171	.000	9.762	2.279	.279	.192	.000	8.301	2.116
16	.478	.106	−.237	2.153	.767	.433	.096	−10.481	.704	−.350
17	.440	.094	−15.048	.330	−1.109	.455	.068	−15.817	.541	−.615
18	.675	.085	−.048	3.136	1.143	.005	2.961	−.224	.695	−.364
19	.609	.130	−.319	1.955	.671	.429	.144	−4.790	.601	−.509
20	.430	.268	15.801	.000	−21.862	.422	.181	.000	9.228	2.222
21	.545	.186	15.156	.000	−20.273	.291	.168	−.010	4.014	1.390
22	.387	.024	−.147	3.334	1.204	.000	.771	−.579	.360	−1.022
23	.523	.107	10.980	.259	−1.352	.523	.134	.000	9.999	2.303
24	1.000	.049	5.261	6.845	1.924	.002	2.814	−.189	.057	−2.870
25	.398	.188	12.575	.085	−2.467	.381	.133	−.018	2.961	1.085
Total	.495 (.198)	.344 (.737)	5.19 (9.0)	2.247 (3.201)	−6.917 (10.802)	.312 (.186)	.424 (.81)	−1.785 (9.425)	2.771 (3.448)	−1.407 (6.215)

Note. Estimated parameters for each subject using the gamma-knowledge reinforcement learning model during High Load and Low Load conditions. Group averages of the estimated parameters are also reported. Group standard deviations are reported in parentheses.

and Low Load conditions. A Wilcoxon signed-ranks test on the average values of overall performance revealed a decrease in the High Load condition ($M = 0.586$, $SD = 0.109$) compared with the Low Load condition ($M = 0.617$, $SD = 0.098$; $z = 2.08$ $p = .036$, $r = .417$), suggesting that loading cognitive control resources made it more difficult for participants to retrieve previously learned information and act strategically. However, in both conditions, all participants scored above a chance level set at 33%. A Wilcoxon signed-ranks test on the average value of participants' overall performance revealed a significant difference between choosing the deck associated with the highest average points during the High Load condition and chance level ($p < 10^{-7}$), and between choosing the deck associated with the highest averaged points during the Low Load condition and chance level ($p <$

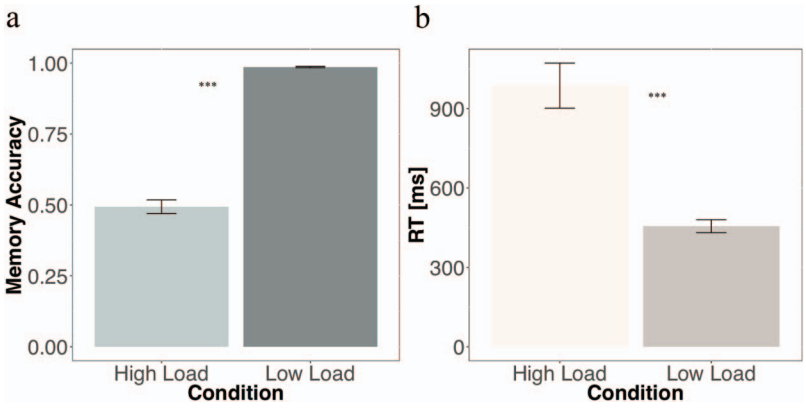

Figure 2. Memory performance and cognitive load manipulation. (a) Memory accuracy measured by averaging trial-by-trial correct memory responses obtained by participants during both High and Low Load conditions. (b) Cognitive load manipulation increased participants' choice reaction time (reaction time [RT] in ms) during the High Load compared with the Low Load condition. Error bars represent the standard error of the mean. *** $p <$ 0.001. See the online article for the color version of this figure.

10^{-7}), indicating that participants played strategically during both load conditions.

Cognitive Control and Decision Strategies

To investigate whether cognitive control plays a role in the resolution of the exploration-exploitation dilemma, we first measured decision strategies when participants selected options unequally during the forced-choice task (unequal information condition) in both the High and Low Load conditions (Figure 3a). We conducted the analysis on the first free-choice trial, being the only trial in which a clear distinction between random and directed exploration can be obtained (Wilson et al., 2014). Trials were classified as *directed exploratory* when participants chose the option that had never been sampled during forced-choice trials, as *exploitative* when participants chose the experienced deck with the highest average of points (regardless of the number of times that deck had been selected during the forced-choice task), and as *random exploratory* when the classification did not meet the previous criteria. The sum of the three strategies defined the total choice probability equal to 1 (choice probability = probability to exploit + probability to random explore + probability to directed explore = 1). We conducted a 2 (condition: High Load, Low Load) × 3 (strategies: exploitation, random exploration, directed exploration) nonparametric ANOVA. The test allows the use of two-way repeated measure ANOVA in a nonparametric setting using aligned rank transformation (e.g., ART package in R; Conover & Iman, 1981). Results showed an effect of strategy, $F(2, 120) = 44.83$, $p < 10^{-15}$, $\eta_p^2 = 0.428$, and a Condition × Strategy interaction, $F(2, 120) = 5.87$, $p = .004$, $\eta_p^2 = 0.089$ (Figure 3a). The effect of condition did not reach the significant threshold ($p = .974$). Post hoc comparisons showed an increase in random exploration in the High Load condition ($M = 0.202$, $SD = 0.123$) compared with the Low Load condition ($M = 0.13$, $SD = 0.09$; $p = .0060$), a decrease in directed exploration in the High Load condition ($M = 0.338$, $SD = 0.177$) compared with the Low Load condition ($M = 0.473$, $SD = 0.198$; $p = .0012$), and an increase in exploitation in the High Load condition ($M = 0.459$, $SD = 0.149$) compared with the Low Load condition ($M = 0.397$, $SD = 0.151$; $p = .031$).

The analysis reported in the previous paragraph appears to suggest that the effect of cognitive load manipulation affected directed and random exploration in an opposite fashion: Directed exploration decreased, whereas random exploration increased, under the High Load compared with the Low Load condition. However, in the unequal information condition, trials labeled as random exploration correspond to the deck of cards that is sampled either 2 or 4 times during the forced-choice task. Therefore, in this condition, trials labeled as random exploration might be confounded with information-based processing (i.e., when subjects select the option observed twice during the forced-choice task). In order to gain insight into this issue we conducted two additional analyses: (a) In the unequal information condition, we repeated the 2 × 3 ANOVA reported in the previous paragraph but only for trials in which random exploratory trials in which those associated with the deck of cards sampled 4 times during the forced-choice task; and (b) we investigated participants' behavior in the equal information condition, in which random exploration was not confounded with the number of observations of each option (being the outcomes of the three options equally experienced; Wilson et al., 2014). In the first analysis, we labeled trials as *exploitative* when the option was associated with highest reward and selected twice during the forced-choice task, *random exploratory* when the option was associated with lowest reward and selected 4 times during the forced-choice task, and *directed exploratory* as previously described. Next, we conducted a 2 (condition: High Load, Low Load) × 3 (strategies: exploitation [2seen]), random exploration [4seen], directed exploration) nonparametric ANOVA. Results showed an effect of strategy, $F(2, 120) = 79.8$, $p < 10^{-15}$, $\eta_p^2 = 0.57$, and a Condition × Strategy interaction $F(2, 120) = 7.48$, $p < 10^{-3}$, $\eta_p^2 = 0.111$, whereas the effect of condition was not signif-

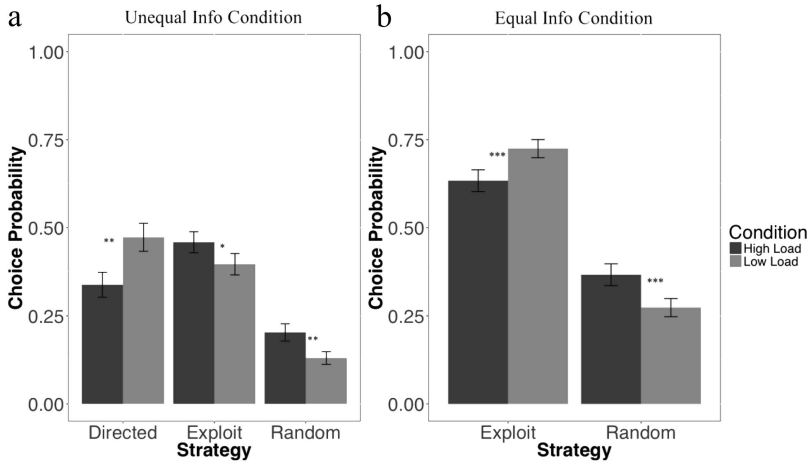

Figure 3. Cognitive load and decision strategies. (a) In the unequal information condition, directed exploration decreased in the High Load condition compared with the Low Load condition, whereas random exploration and exploitation showed the opposite trend. (b) In the equal information condition, random exploration increased under the High Load condition, whereas exploitation decreased. Error bars represent the standard error of the mean. Info = information. * $p < 0.05$. ** $p < 0.01$. *** $p < 0.001$.

icant ($p = .137$). Post hoc comparison revealed an increase in random exploration in the High Load ($M = 0.119$, $SD = 0.073$) compared with the Low Load ($M = 0.08$, $SD = 0.053$) condition ($p = .025$), whereas exploitation did not differ. Results concerning directed exploration are already reported in the previous paragraph. In the second analysis, we investigated the effect of cognitive load manipulation on decision strategies when participants were forced to equally select options (equal information condition; Figure 3b). We classified choices as *exploitative* when participants chose the experienced deck with the highest average of points, and *random explorative* otherwise. A 2 (condition: High Load, Low Load) × 2 (strategy: exploitation, random exploration) nonparametric ANOVA on participants' choices showed an effect of strategy, $F(1, 45) = 64.06$, $p < 10^{-10}$, $\eta_p^2 = 0.587$, and a Condition × Strategy interaction, $F(1, 45) = 5.9$, $p = .019$, $\eta_p^2 = 0.116$. Post hoc comparisons revealed an increase in random exploration in the High Load condition ($M = 0.366$, $SD = 0.155$) compared with the Low Load condition ($M = 0.273$, $SD = 0.129$; $p = .0009$), and a decrease in exploitation in the High Load condition ($M = 0.633$, $SD = 0.155$) compared with the Low Load condition ($M = 0.725$, $SD = 0.129$; $p = .001$). Taken together, these analyses confirm that cognitive control manipulation affected the two exploratory strategies in a different fashion.

Subsequently, we investigated whether the results reported in the previous paragraph could have been driven by an ineffective High Load manipulation in trials in which participants incorrectly performed the working memory task. To do so, we compared RTs from correct and incorrect memory trials during the High Load condition. If the behavioral pattern observed in the previous analysis was driven by an ineffective load manipulation during incorrect memory trials, participants should have shown differences in their RTs as a function of memory accuracy. We computed participants' RTs during correct and incorrect memory trials and compared the average values. A Wilcoxon signed-ranks test on choice RTs showed no differences between correct ($M = 1{,}023$ ms, $SD = 562$ ms) and incorrect ($M = 993$ ms, $SD = 423$ ms) memory trials in all free-choice trials ($z = 0.04$, $p = .979$, $r = .008$) and a marginal difference in the first free-choice trials (correct: $M = 2{,}036.8$ ms, $SD = 1{,}831.6$ ms; incorrect: $M = 2{,}116.3$ ms, $SD = 1{,}425.2$ ms; $z = -1.95$, $p = .051$, $r = -.39$). However, this marginal difference was in the direction of higher RTs for incorrect trials, as participants were taking more time to retrieve incorrectly memorized sequence. Overall, these results suggest that even if participants were not correctly performing the memory task, they were still in a "loaded state" during the High Load condition, suggesting that the observed effects on the decision strategies were a direct consequence of lowering cognitive control resources.

Randomness Versus Information Integration Under Cognitive Load

Our previous analysis showed that manipulating cognitive control resources affected how participants balanced the exploration-exploitation dilemma, exploring more randomly overall during high-working-memory-load conditions. In this section, we asked whether this effect was related to an increase in the randomness in participants' choices or to alterations in reward and information processing that subtend the resolution of the dilemma through directed exploration (Cogliati Dezza et al., 2017). To better investigate the mechanisms affected by the load manipulation, we fit the gkRL model to all participants' first free choices during both the High and Low Load conditions to obtain the estimates of the values of the following parameters: learning rate α, inverse of the temperature β, the nonlinear parameter γ, and the information parameter ω (see Table 1). We then compared the estimated parameters for the High Load condition with the parameters of the Low Load condition to investigate the effect of the cognitive control manipulation. As expected, because the learning processes during the forced-choice task were not affected, a Wilcoxon signed-ranks test on the learning rate α showed no difference between the Low Load ($M = 0.426$, $SD = 0.249$) and High Load ($M = 0.456$, $SD = 0.3$) conditions ($z = -0.4171$, $p = .691$, $r = -.083$). Furthermore, a Wilcoxon signed-ranks test on the inverse temperature parameter β showed no difference between the Low Load ($M = 0.606$, $SD = 1.383$) and High Load ($M = 0.865$, $SD = 1.8$) conditions ($z = 0.094$, $p = .937$, $r = .019$). Additionally, a Wilcoxon signed-ranks test on the parameter γ showed no difference between the Low Load ($M = 1.66$, $SD = 3.44$) and High Load ($M = 1.44$, $SD = 2.62$) conditions ($z = -0.444$, $p = .672$, $r = -.089$). On the contrary, the information parameter ω showed a decrease in the High Load condition ($M = -2.81$, $SD = 51.76$) compared with the Low Load condition ($M = 4.82$, $SD = 9.89$; $z = -2.058$, $p = .039$, $r = -.412$), suggesting that the increase in random exploration was related to an inability to integrate the learned information into a choice value rather than an increase in the randomness of participants' choices or by alteration in how information is decay with time (Figure 4a).

Furthermore, we fitted the model to all free-choice trials so as to have a more comprehensive view over the underlying process as well as to obtain a better estimate of the parameter values because of the higher number of data points (see Table 2). As before, a Wilcoxon signed-ranks test showed no difference between the Low Load and High Load conditions, neither for the inverse temperature parameter β ($M = 0.344$, $SD = 0.737$ vs. $M = 0.424$, $SD = 0.81$; $z = 0.202$, $p = .853$, $r = .04$) nor for the γ parameter ($M = 2.247$, $SD = 3.201$ vs. $M = 2.771$, $SD = 3.448$; $z = -0.336$, $p = .751$, $r = -.067$). Again, a Wilcoxon signed-ranks test on the information parameter ω did reveal a decrease in information integration from the low load ($M = 5.19$, $SD = 9$) to the High Load ($M = -1.993$, $SD = 9.3$; $z = -3.35$, $p = .0003$, $r = -.67$). However, the same test applied to the learning rate α revealed a decrease in the speed of integration of new reward information from Low Load ($M = 0.495$, $SD = 0.198$) to High Load ($M = 0.312$, $SD = 0.186$; $z = 3.108$, $p = .001$, $r = -.621$; Figure 4b). The effect on learning rate in this analysis is explained by the fact that we considered all free-choice trials during which participants were performing the memory task while repeatedly selecting options. As a consequence, the ability to integrate new reward information (expressed by the learning rate) was also affected.

As an additional check, we fit the gkRL model exclusively on the free-choices trials in which memory responses were correct in both the low- and High Load conditions. Wilcoxon signed-ranks test confirmed our previous results: no differences in parameter β ($z = -0.525$, $p = .615$, $r = -.105$) and parameter γ between low- and High Load conditions ($z = -1.0$, $p = .325$, $r = -.202$), whereas a higher α was observed in the Low Load

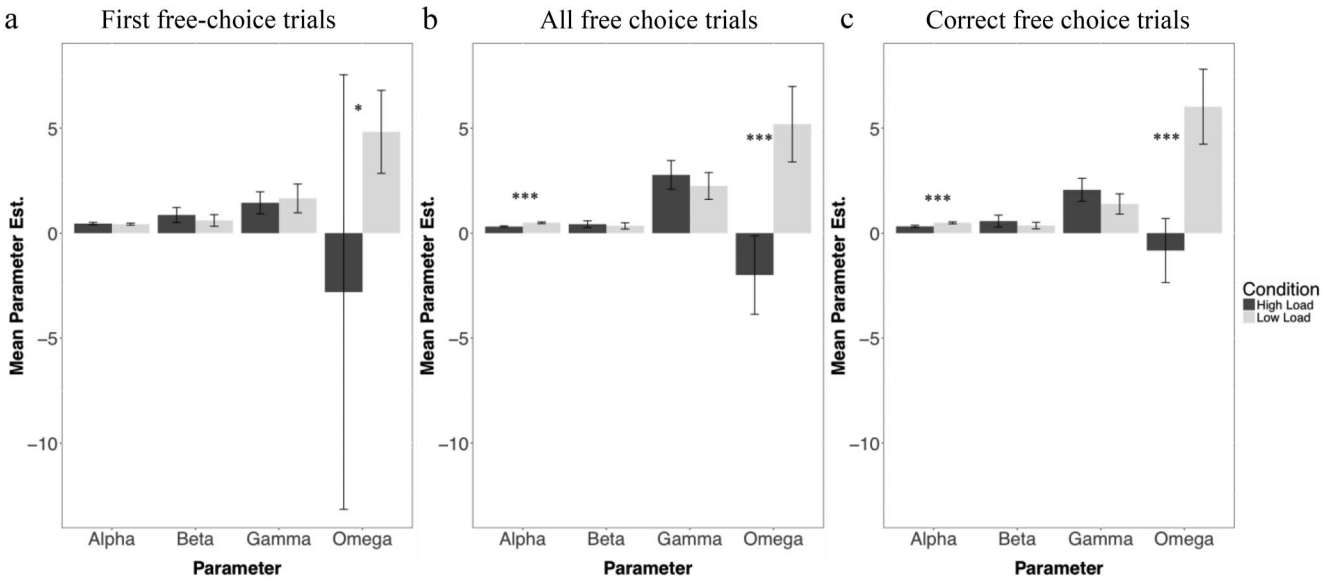

Figure 4. Information integration. (a) First free trials: Model fit on the first free choice only revealed a decrease in the information weigh parameter omega (ω - that modulates to integration of information into choice values) during the High Load compared with the Low Load condition, whereas the inverse of temperature beta (β), the learning rate alpha (α), and the gamma (γ) parameter were not affected by the cognitive load. (b) All-free trials: Model fit on all free choices showed a decrease in information parameter ω and the learning rate α in the High Load condition, whereas both β and γ were not affected by the cognitive manipulation. (c) Correct memory choices: Model fit on the trials where participants correctly performed the memory task. The results showed the same pattern observed when fitting all free choices. Error bars represent the standard error of the mean. * $p < 0.05$. *** $p < 0.001$.

($M = 0.493$, $SD = 0.229$) compared with the High Load ($M = 0.321$, $SD = 0.248$) condition ($z = 2.516$, $p = .001$, $r = .503$). A higher information parameter ω was also obtained in the Low Load ($M = 6.02$, $SD = 8.9$) compared with the High Load ($M = -0.827$, $SD = 7.63$) condition ($z = 3.4$, $p = .0002$, $r = .686$; Figure 4c).

Cognitive Control and Information Integration

Following the results reported in the previous paragraph, cognitive load seems to affect participants' ability to integrate learned information into choice values in order to solve exploration-exploitation problems. As a further investigation, we asked whether a standard RL (sRL) model that learned reward values following Equation 1 and entered directly in Equation 4 without any integration of information, could better explain this "inability" in integrating information during the cognitive control manipulation. To do so, we compared fits of both the gkRL and sRL models. During the fitting procedure, we computed negative-log likelihoods of both models and their model evidence (or the log model evidence—the probability of obtaining the observed data given a particular model). We adopted an approximation to the (log) model evidence, namely, the Bayesian information criterion (BIC; Schwarz, 1978). We conducted a frequentist analysis with BIC values of the two models (fitted to the first free-choice trials) entered into a t test. Results showed that during the Low Load condition the gkRL model ($BIC_{gkRL} = 184$) best represented participants' data compared with sRL ($BIC_{sRL} = 203$), $t(24) = -3.034$, $p = .005$, $d = 0.455$, replicating our previous findings on reward and information integration during this new version of the bandit task (Cogliati Dezza et al., 2017). However, in the High Load condition, neither the gkRL model ($BIC_{gkRL} = 218$) nor the sRL model ($BIC_{sRL} = 214$) better represented participants' data, $t(24) = 0.437$, $p = .666$, $d = -0.076$. To better understand this point, we individually investigated the BIC values of each model (see Figure 5). Although in the Low Load condition, the performance of the majority of participants was better explained by the gkRL model (Figure 5a), in the High Load condition, approximately 60% of participants were better represented by the sRL model (whereas the behavior of 20% were better explained by the gkRL model, and 20% were equally explained by both models; Figure 5b), confirming that during the High Load condition, information processing was heavily compromised, and that for the majority of subjects, the computation of information was nullified. Furthermore, we extended the comparison of the two computational models to all free-choice trials to have an exhaustive understanding of the hidden processes. Contrary to our previous model comparison in the High Load condition, results showed that, when fit to all free-choice trials, the gkRL model ($BIC_{gkRL} = 802$) best represented participants' data compared with the sRL model ($BIC_{sRL} = 849$), $t(24) = -3.4$, $p = .002$, $d = 0.258$ (we obtained the same result in the Low Load condition, so the results are not reported here).

A possible reason behind this apparently incoherent result could be related to the working memory process itself. The memory

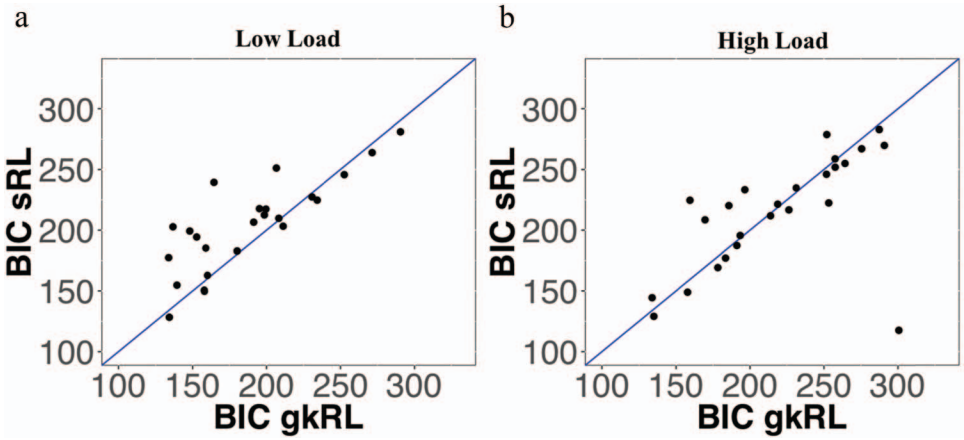

Figure 5. Comparative fit of the gkRL and sRL. The comparison of the fit is based on the Bayesian information criterion (BIC) values of both models during the (a) Low Load and (b) High Load conditions. Each point represents one participant. The sRL fit better when the point is below the identity line. When a point lays on the identity line, the models equally explain participants' behavior. See the online article for the color version of this figure.

sequence was presented to participants at the beginning of the free-choice task only: Cognitive load may be reduced during later free-choice trials either as a consequence of inability to maintain the complete sequence over the course of the free-choice task (and thus freeing cognitive resources for making choices), or because cognitive demands related to maintaining the sequence are higher immediately following the presentation of the sequence. We therefore investigated participants' behavior during all free-choice trials to better clarify this point. However, after the first free-choice trial, it is not possible to distinguish between random and directed exploration due to a confound between reward and information (Wilson et al., 2014). For this reason, in order to investigate participants' behavior during the all-free-choice trials, we focused on information-based processes only. To do so, we computed the probability of selecting the least-seen deck (the option visited the least number of times in the previous trials), the most-seen deck (the option visited the most number of times in the previous trials), and the middle-seen deck (when previous criteria did not match) during both load conditions. When we investigated the behavior globally, the analysis gave similar results observed in the previous behavioral analysis (see the Cognitive Control And Decision Strategies section), in which the probability of selecting the least-seen option was reduced during the High Load ($M = 0.255$, $SD = 0.099$) compared with the Low Load ($M = 0.304$, $SD = 0.071$) condition ($z = -2.652$, $p = .006$, $r = -.53$), whereas the most-seen showed the opposite pattern an increase in the High Load ($M = 0.573$, $SD = 0.121$) compared with the Low Load ($M = 0.531$, $SD = 0.092$) condition ($z = 2.18$, $p = .028$, $r = .436$; the probability of choosing the middle-seen option did not differ, and so we will not consider this strategy any further; Figure 6a). However, investigating the trial-by-trial probability revealed a more exhaustive view. Indeed, the above result was true only for the first three free-choice trials (all $ps < 10^{-2}$), whereas we did not observe differences in terms of the most-seen and least-seen options during the last three trials (all $ps > .05$; Figure 6b). These results suggest that the effect of cognitive load was greatest during the first free-choice trials and vanished during the last trials, suggesting that the reason behind the better performance of gKRL compared with sRL in explaining all participants' free choices was related to a decrease in cognitive load in the last trials of each game. Considering that the previous analyses focused on information only, it is possible that additional factors may inform choice behavior in free-choice trials. To examine this, we also computed switch/stay probabilities for free-choice trials. Switch/stay behavior changed in the High Load ($M_{switch} = 0.416$, $SD_{switch} = 0.165$; $M_{stay} = 0.584$, $SD_{stay} = 0.165$) compared with Low Load ($M_{switch} = 0.476$, $SD_{switch} = 0.118$; $M_{stay} = 0.533$, $SD_{stay} = 0.118$) condition (both $ps = .042$). However, differences in switch/stay behavior were most apparent on the first free-choice trial—subjects tended to switch choices but did so more often in the Low Load condition (see Figure 7). Results showed that in the last trials of each game, stay (switch) probability did not change between High Load and Low Load conditions (all $ps > .05$), confirming that a decrease in cognitive load occurred in the last trials of each game.

From the Model to Behavior

The results reported in the previous paragraph demonstrate that the gkRL model better accounts for our behavioral data relative to sRL. In order to demonstrate that the gkRL model parameters are behaviorally relevant, we correlated the differences observed between the two load conditions in the information-integration parameter ω with the differences in exploitation in the unequal information condition. If the model captures behavioral dynamics, we should expect increased differences between the estimate of parameter ω in the two load conditions as well as increased differences in exploitation between the two load conditions. We observed a positive correlation between the difference in ω and the difference in exploitation (Pearson correlation, $r[23] = .413$, $p = .039$), suggesting that reduction of the integration of new information was associated with increased exploitative behaviors. Addi-

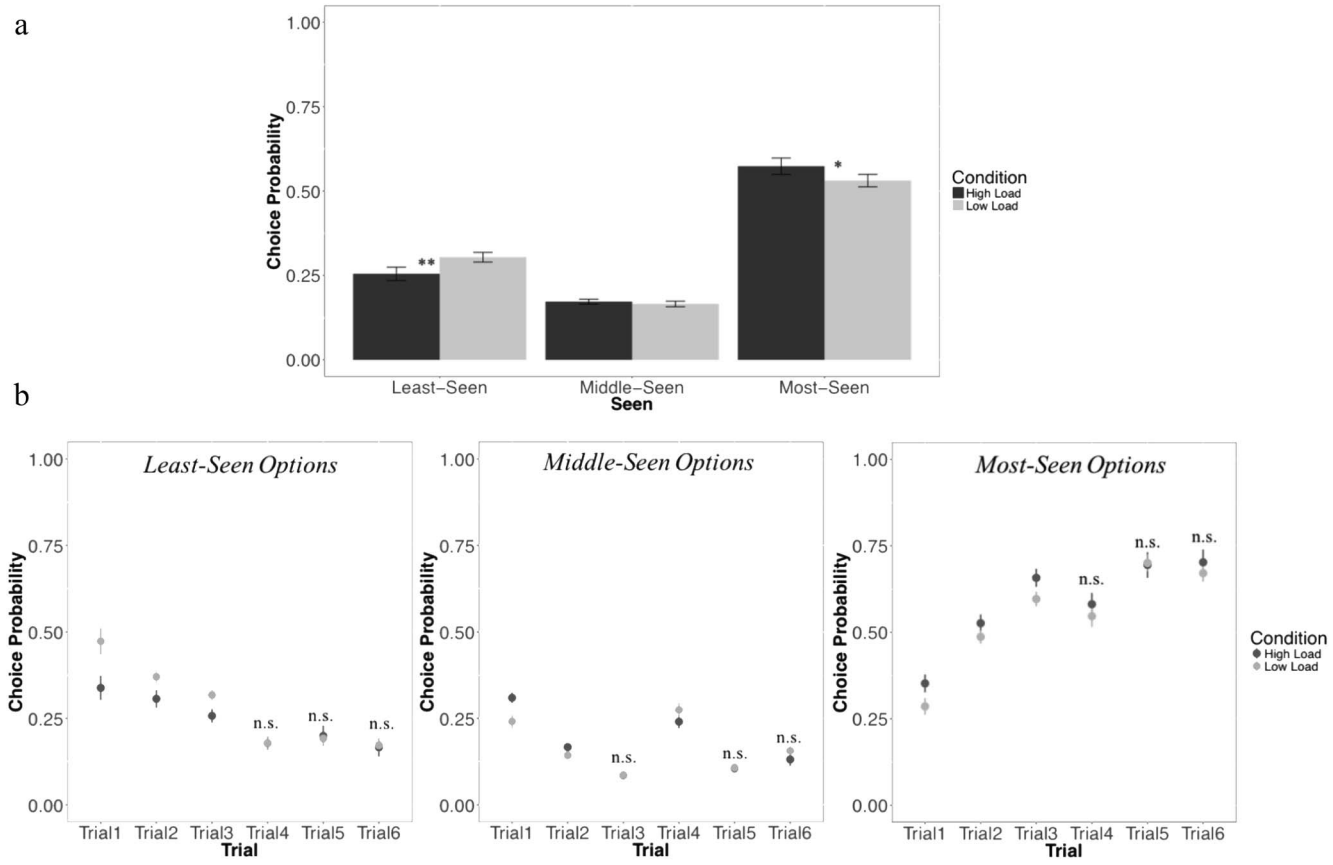

Figure 6. Seen analysis. (a) Probability to choose the option seen least, middle, and most of the time during the free-choice task. Choices toward the least-seen option decreased during the High Load compared with the Low Load condition, whereas choices toward most-seen options showed the opposite pattern. (b) Probability to choose the option seen most (most-seen options), least (least-seen options), and middle (middle-seen options) of the time during the free-choice task split by trial. During the first three free trials, the probability to choose the least-seen option (and most-seen option) differs significantly, whereas in the last free-choice trials, no difference was observed. To avoid overloading the visualization, we reported only when the comparisons did not reach significance threshold. Error bars represent the standard error of the mean. * $p < 0.05$. ** $p < 0.01$.

tionally, simulations of the model are also able to reproduce the condition-dependent behavioral results we observe in our data. We simulated the gkRL model 80 times under the two loading conditions. In the High Load condition, ω values were randomly drawn from a uniform distribution with mean −2, whereas for the Low Load condition, the mean was set to 5. The other parameters did not change between the two conditions, and their values were randomly chosen from a uniform distribution, with the mean set around the mean values observed in participants' data. We then labeled the model's choices in the unequal information condition as *directed exploratory*, *random exploratory*, and *exploitative*. We conducted a 2 (condition: High Load, Low Load) × 3 (strategies: exploitation, random exploration, directed exploration) nonparametric ANOVA. Results showed an effect of strategy, $F(2, 395) = 223.04$), $p < 10^{-15}$, $\eta_p^2 = 0.53$, and a Condition × Strategy interaction, $F(2, 395) = 240.52$, $p < 10^{-15}$, $\eta_p^2 = 0.549$. The effect of condition did not reach the significant threshold ($p = .5$). The results mimicked the same behavioral pattern observed in participants' data (Figure 8a). Additionally, we computed random exploration and exploitation in the equal information condition. We conducted a 2 (condition: High Load, Low Load) × 2 (strategies: exploitation, random exploration) nonparametric ANOVA. Results showed an effect of strategy, $F(2, 237) = 382.89$, $p < 10^{-15}$, $\eta_p^2 = 0.617$; however, neither an effect of Condition × Strategy nor of condition was observed (all ps $> .05$; Figure 8b). The behavior of the model in the equal information condition, however, did not replicate the findings observed in participants' data. We discuss this result in more depth in the next section.

Cognitive Control and Value Degradation

In order to understand the underlying mechanisms affected in the equal information condition that are not captured by the information-integration account expressed by the gkRL model, we implemented a new version of the gkRL model—the value gamma-knowledge RL (vgkRL). The rationale behind this additional model implementation is that cognitive load might have affected processes concerning both information integration (as

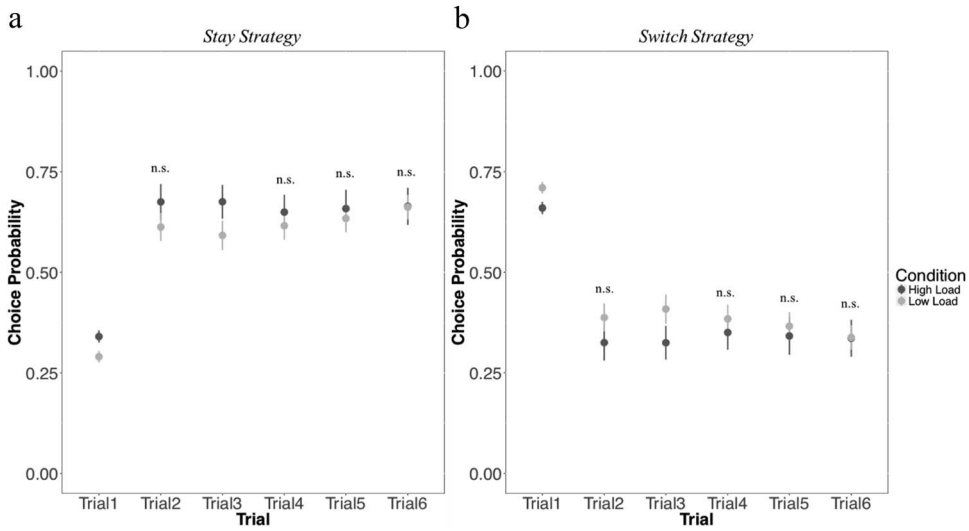

Figure 7. Switch/stay strategy. (a) Probability of staying with the same option chosen at trial t–1 during the free-choice task. During last free-choice trials, the probability to stay with the same option did not differ between the two loading conditions. (b) Probability of switching from the option chosen at trial t–1 during the free-choice task. During last free-choice trials, the probability to switch did not differ between the two loading conditions. Error bars represent the standard error of the mean. n.s. = not significant.

captured by the gkRL model) as well as reward information. Indeed, the gkRL model was developed primarily in order to capture participants' behavior in the unequal sampling scenario, in which differences in information are expected to have a large influence on exploration-exploitation decisions (Cogliati Dezza et al., 2017). However, model simulations in the equal information condition appear to suggest that cognitive load may additionally degrade the integration of reward information into an overall choice value. In order to investigate this reward degradation account, the vgkRL adds an integration of reward values on top of the information integration expressed in gkRL. Equation 3 thus becomes

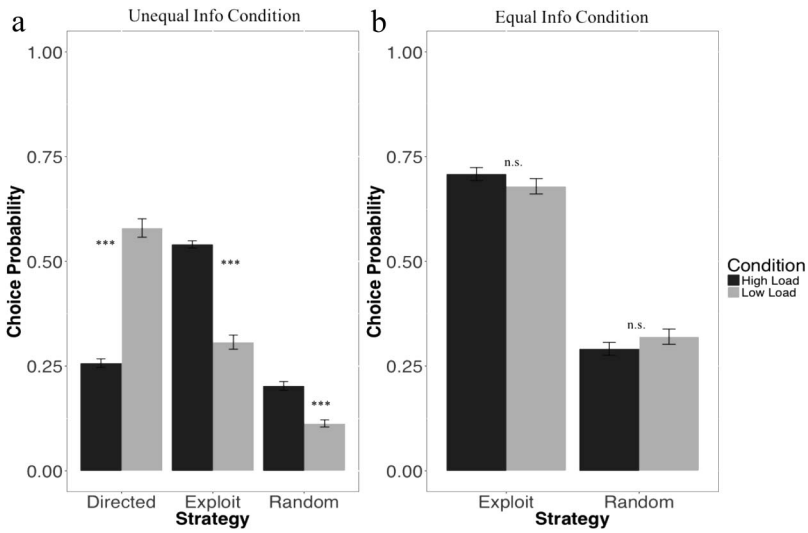

Figure 8. Gamma-knowledge reinforcement learning model simulation. (a) In the unequal information condition, the model simulated under the two loading conditions reproduced the same behavioral pattern observed in participants: Directed exploration decreased in the High Load condition, whereas random exploration and exploitation increased in Low Load condition. (b) In the equal information condition, no behavioral differences in exploitation and random exploration were observed between the two loading conditions. Only the comparisons that did not reach significance threshold are reported. Info = information. *** $p < 0.001$. n.s. = non significant.

$$V_{t,j}(c) = (Q_{t+1,\,j}(c) \times \rho) - (I_{t,j}(c) \times \omega), \quad (5)$$

where ρ indicates the degree by which reward values are integrated in choice values. We fitted vgkRL to participants' data and simulated the model using the retrieved parameters. We then analyzed model behavior in both unequal and equal information conditions. In the unequal condition, we conducted a 2 (condition: High Load, Low Load) × 3 (strategies: exploitation, random exploration, directed exploration) nonparametric ANOVA. Results showed an effect of strategy, $F(2, 110) = 21)$, $p < 10^{-7}$, $\eta_p^2 = 0.144$, and a Condition × Strategy interaction, $F(2, 110) = 4.79$, $p = .01$, $\eta_p^2 = 0.743$ (Figure 9a). The effect of condition did not reach the significant threshold ($p = .9$). Post hoc comparisons revealed the same pattern observed in participants' behavior in which directed exploration decreases, whereas random exploration increases, in the High Load compared with the Low Load condition (all ps $< .05$). On the contrary, exploitation did not differ between the two conditions ($p > .05$). Subsequently, we conducted a 2 (condition: High Load, Low Load) × 2 (strategies: exploitation, random exploration) nonparametric ANOVA in the equal information condition. Results showed an effect of strategy, $F(2, 72) = 23.87)$, $p < 10^{-5}$, $\eta_p^2 = 0.249$, and a Condition × Strategy interaction, $F(2, 72) = 6.16$, $p = .015$, $\eta_p^2 = 0.079$ (Figure 9b). The effect of condition did not reach the significant threshold ($p = .986$). Post hoc comparisons revealed the same pattern observed in participants' behavior in which exploitation decreases, whereas random exploration increases, in the High Load compared with the Low Load condition (all ps $< .05$). These results seem to suggest that on the top of the information degradation process occurring in the unequal information condition, cognitive load also affected reward value degradation captured by the ρ parameter in vgkRL model. Therefore, cognitive load appears to specifically interfere with the ability to combine reward and information in order to inform choices. To better test this hypothesis, we compared the estimated parameters of the model between the two conditions. Unfortunately, the analysis did not reveal any differences in the estimated parameters between the Low Load and High Load conditions (all ps $> .05$). The reason behind this counterintuitive result might be that when adding parameters to the model, higher numbers of data points are necessary in order to obtain a reliable estimate within the same statistical power. Thus, the fitting procedure was less powerful and less able to recover the accurate estimates.

Discussion

The results of this study challenge a popular view concerning the cognitive mechanisms underlying the resolution of the exploration-exploitation dilemma. Specifically, following this perspective, the dilemma is considered as a binary process and cognitive control as the main underlying mechanism that is required in order to override default exploitative strategies in favor of exploration of the surrounding environment (Cohen et al., 2007; Daw et al., 2006). Our results showed that, indeed, the need for cognitive control seems necessary when resolving the dilemma. However, increased cognitive load appears to affect only one aspect of exploration, namely, directed exploration, and the effect of cognitive load on exploration seems to be driven mostly by information degradation. Additionally, our results unveiled a different facet of exploitative behaviors that moves away from the traditional view of exploitation as a "default strategy." Together, these findings shed additional light on the mechanisms underlying adaptive control and behavior and suggest new approaches for interpreting the exploration-exploitation dilemma. In the following, we discuss the implications of our main results.

In line with what could be expected because of dual-task interference (Herath, Klingberg, Young, Amunts, & Roland, 2001), participants' choice RTs were affected by high cognitive load, suggesting that participants cognitive control resources were effectively reduced in this condition. Further analyses showed that high cognitive load affected participants' performance on the gam-

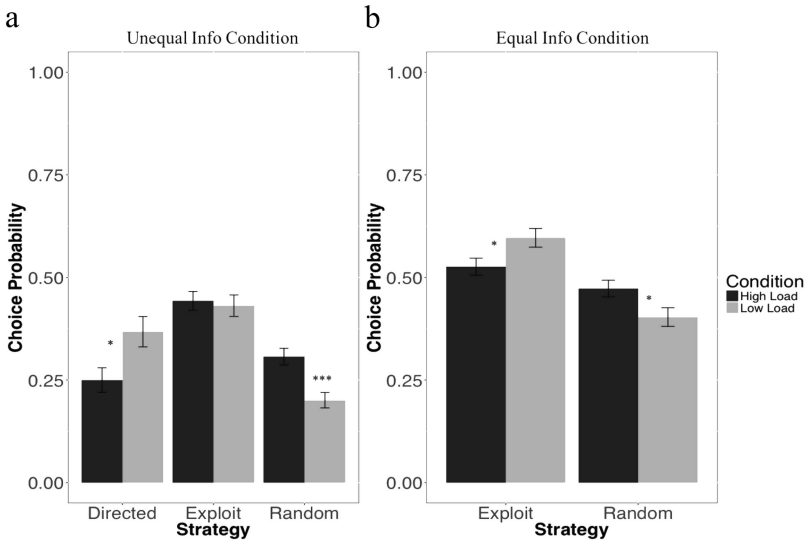

Figure 9. Value gamma-knowledge reinforcement learning simulation. This simulation reproduced a similar pattern observed in participants' data in both the (a) unequal information condition, and (b) equal information condition. Info = information. * $p < 0.05$. *** $p < 0.001$.

bling task in terms of choosing the option associated with highest reward (i.e., overall performance). Under both load conditions, overall performance was above chance-level. However, during the High Load condition, participants were slower in integrating new evidence, as shown by the decrease in the learning rate during the free-choice task, which, in turn, might explain the decrease in overall participants' performance.

One of the main results of this study concerns the antagonist effects of cognitive load on the two exploratory strategies. Specifically, increased cognitive load resulted in a decrease in directed exploration and an increase in random exploration, suggesting that directed exploration depends on the availability of sufficient control resources, and that depletion of such resources promotes random exploration. This result presents a different picture concerning the involvement of cognitive control in the resolution of the exploration-exploitation dilemma compared with that suggested by the behavioral control hypothesis (Cohen et al., 2007; Daw et al., 2006). Resolving the dilemma through exploration seems to not be a unitary process that always requires cognitive resources to be mustered, independent of the type of exploratory strategies adopted. On the contrary, the resolution of the dilemma through exploration is a multifaceted phenomenon (Somerville et al., 2017; Warren et al., 2017; Wilson et al., 2014; Zajkowski et al., 2017), and cognitive control seems to intervene only when exploring the environment in a directed, intentional manner. These results are in line with recent studies that suggest that random and directed exploration are distinct strategies, even if based on a common exploratory drive (Cogliati Dezza et al., 2017; Zajkowski et al., 2017).

Furthermore, when interfering with the resolution of the dilemma, cognitive control cooperates with those aspects of exploration related with the integration of information into choice values. Under cognitive load, participants were more prone to stay with the same option (as shown by effects of cognitive load in the switch/stay behavior), penalizing the search for new information. This result is in line with several studies on information-based processes concerning the exploration-exploitation dilemma that collectively highlight a tight association between information-based exploration (directed exploration) with prefrontal areas involved in higher level cognitive processes (Badre et al., 2012; Cavanagh et al., 2012) as well as the prefrontal dopamine network (Frank et al., 2009; Kayser, Mitchell, Weinstein, & Frank, 2015). However, our results appear to contrast with a study by Daw and colleagues (2006) that suggested a crucial role for top-down control processes in random exploration. In their study, activity in brain regions associated with higher level cognitive functions (i.e., frontopolar cortex) was associated with the probability of randomly exploring options. Frontopolar cortex was subsequently associated with switching between strategies instead of targeting exploratory strategy itself (Boorman, Behrens, Woolrich, & Rushworth, 2009), and transcranial magnetic stimulation studies of this region affected only directed exploration (Zajkowski et al., 2017). Clearly, more research is needed to understand the neuronal and neurochemical mechanisms underlying exploration in light of the new and recent evidence on directed and random exploration.

Our results are in line with a recent finding that showed higher cognitive costs associated with those processes involved in reflexive exploration (Otto et al., 2014). Specifically, cognitive load seems to affect participants' ability to use a model of the environment in which environmental statistics (i.e., state transition probabilities) and reward structure are integrated into choice values in order to guide exploratory behaviors. However, our results suggest a more nuanced view concerning this phenomenon: The results of our model fits suggested that reducing cognitive resources specifically affected those processes involved in information integration, whereas processes involved in transforming probability distributions into action selection (decreasing or increasing the level of noise in the system through a softmax function) were unaffected. Moreover, the effect of cognitive load on information is restricted to integration and not to other aspects of the information processing, such as information decay (which might be captured by differences in the gamma parameter). In our study, however, we approached the computational problem using a model-free strategy, in which choices are only driven by past experience (information and reward history) without a representational characterization of the environment (contrary to a model-based strategy, in which choices are driven by the model of the world; Daw, Niv, & Dayan, 2005). It might be possible that in real-life scenarios, humans adopt model-based approaches when facing exploration and exploitation problems, requiring more complex and resource-intensive computations that are only approximated by the manner in which information is integrated in the gkRL model. The relation between model-based strategy and information integration should be addressed by future research.

Our results further question the interpretation of exploitation as a default strategy that requires cognitive control to be inhibited (Cohen et al., 2007; Daw et al., 2006). Contrary to what might have been expected by the behavioral control hypothesis (Cohen et al., 2007), exploitation was affected by the cognitive control manipulation in such a way that when participants visited the options the same number of times (i.e., equal information condition), they decreased exploitative choices during the High Load condition. This finding seems to suggest that, under certain scenarios, cognitive control is necessary to achieve exploitation, as in the other goal-directed behaviors. That is, choosing to exploit requires cognitive resources in a fashion similar to choosing to explore. Our results are in line with recent findings on cognitive foraging, in which exploring other patches or exploiting familiar patches involved similar cognitive mechanisms (Hills, Todd, & Goldstone, 2010). Our results also provide support for the view that considers exploitation not only as the strategy that selects best rewarded actions but also as a strategy that relies on cognitive control resources to maintain task demands (Hills et al., 2010). Sticking with the same option can be considered as a subgoal of the higher goal of maximizing reward in the long run, and maintaining attention between competing task demands required higher cognitive control functions (e.g., the cocktail party phenomenon; Conway, Cowan, & Bunting, 2001; Hills et al., 2010). A drawback, however, is that our model was unable to capture this phenomenon (Figure 8b). Indeed, gkRL was developed in order to capture human behavior in unequal sampling scenarios (Cogliati Dezza et al., 2017). In order to understand the underlying mechanisms of the effect of cognitive load on the exploitation, we presented an implementation of the gkRL model in which the integration of reward into choice value was also modulated. Simulations of this model showed that the reward value degradation might be the underlying mechanism behind the decrease in exploitation in the equal information condition. However, the limited number of trials

available in our paradigm precluded a definitive answer to this question. Further work is needed in order to understand how cognitive control might influence choice value computation.

Taken together, our results suggest a new perspective on the exploration-exploitation dilemma as the product of multiple competing control modes that jointly promote adaptive behavior through increased emphasis on stability or flexibility. Similar to cognitive search modes (Hommel, 2012), the differences between these control modes might be in the control style they call for: a divergent decision-making style—one goal representation that diverges to different action selections (or perceptual representations in the case of cognitive search)—and a convergent style, in which a potential number of possible actions (or a number of representations) converges toward an optimal solution (Hommel, 2012). At the neural level, these different modes may be represented by tonic and phasic activity in the locus coeruleus (LC) expressed by the release of norepinephrine (NE; Aston-Jones & Cohen, 2005). LC-NE is the target of projections from cortical regions implicated in cognitive control and adaptive behavior, including regions involved in processing information regarding behaviorally salient changes in the environment (e.g., anterior cingulate cortex. anterior insula, and orbitofrontal cortex). Following unexpected changes in the environment, tonic LC-NE activity may favor adaptive exploration by allowing disengagement from current task demands (Yu & Dayan, 2005). On the other hand, in stable environments, phasic LC-NE activity may promote exploitative behavior by increasing attention toward task-relevant stimuli and maintenance of the current goal (Aston-Jones & Cohen, 2005; Jepma & Nieuwenhuis, 2011). This perspective, however, leaves many questions unanswered. For example, the interaction between these control modes and the regions previously associated with exploration (i.e., frontopolar cortex) is still unknown and needs to be addressed by future research. Moreover, random exploration, but not directed exploration, was affected by pharmacological manipulation of baseline NE levels (Warren et al., 2017), questioning how the LC-NE system may control the two exploratory strategies and which is the role of random exploration in this mode-based trade-off. So far, random exploration seems to be a low-level (Warren et al., 2017) or automatic action control process (Humphries et al., 2012) that might be necessary when a less engaging or faster behavioral adaption is required. However, the exact manner in which low-level control interacts with higher cognitive control remains an open question and should be the subject of future research.

Although our study adds additional perspective on the cognitive mechanisms underlying the resolution of the exploration-exploitation dilemma by humans, there are nonetheless limitations that may influence the scope of our results. Besides the limitation of the computational model previously discussed, the absence of a horizon manipulation in our paradigm makes it impossible to distinguish whether the increase in random exploration in the High Load condition was related to random exploration itself (changes in randomness in long horizon) or by an overall increase in randomness (Krueger, 2017). In the same line, the information-integration parameter was not horizon-dependent. Thus, we cannot explain the effect of cognitive load on information integration on a trial basis. Additionally, although ambiguity appears to modulate the tension between exploration and exploitation (Krueger, 2017; Wilson et al., 2014), we did not specifically investigate this aspect in this study. Lastly, we did not compute participants' memory span, preventing us from delineating individual profiles concerning the efficacy of our experimental manipulation.

Regardless of these limitations, using a recently developed behavioral paradigm (Wilson et al., 2014), we disentangled the role of cognitive control in the resolution of the exploration-exploitation dilemma. Our results emphasized the multifaceted nature of the resolution of the dilemma and suggest that multiple cognitive control modes are the underlying cognitive mechanisms. This study is in line with a new perspective on how to look at the exploration-exploitation dilemma, and provides a formal foundation within which to explore pathologies of goal-directed behavior such as manifest in addiction, obsessive–compulsive disorders, and attentional deficits.

References

Aston-Jones, G., & Cohen, J. D. (2005). Adaptive gain and the role of the locus coeruleus-norepinephrine system in optimal performance. *The Journal of Comparative Neurology, 493,* 99–110. http://dx.doi.org/10.1002/cne.20723

Baddeley, A., Emslie, H., Kolodny, J., & Duncan, J. (1998). Random generation and the executive control of working memory. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 51,* 819–852. http://dx.doi.org/10.1080/713755788

Badre, D., Doll, B. B., Long, N. M., & Frank, M. J. (2012). Rostrolateral prefrontal cortex and individual differences in uncertainty-driven exploration. *Neuron, 73,* 595–607. http://dx.doi.org/10.1016/j.neuron.2011.12.025

Boorman, E. D., Behrens, T. E., Woolrich, M. W., & Rushworth, M. F. (2009). How green is the grass on the other side? Frontopolar cortex and the evidence in favor of alternative courses of action. *Neuron, 62,* 733–743. http://dx.doi.org/10.1016/j.neuron.2009.05.014

Cavanagh, J. F., Figueroa, C. M., Cohen, M. X., & Frank, M. J. (2012). Frontal theta reflects uncertainty and unexpectedness during exploration and exploitation. *Cerebral Cortex, 22,* 2575–2586. http://dx.doi.org/10.1093/cercor/bhr332

Cogliati Dezza, I., Yu, A. J., Cleeremans, A., & Alexander, W. (2017). Learning the value of information and reward over time when solving exploration-exploitation problems. *Scientific Reports, 7,* 16919. http://dx.doi.org/10.1038/s41598-017-17237-w

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences, 362,* 933–942. http://dx.doi.org/10.1098/rstb.2007.2098

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician, 35,* 124–129.

Conway, A. R., Cowan, N., & Bunting, M. F. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic Bulletin & Review, 8,* 331–335. http://dx.doi.org/10.3758/BF03196169

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

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature, 441,* 876–879. http://dx.doi.org/10.1038/nature04766

D'Esposito, M., Postle, B. R., Ballard, D., & Lease, J. (1999). Maintenance versus manipulation of information held in working memory: An event-

related fMRI study. *Brain and Cognition, 41,* 66–86. http://dx.doi.org/10.1006/brcg.1999.1096

Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience, 12,* 1062–1068. http://dx.doi.org/10.1038/nn.2342

Gittins, J., & Jones, D. (1974). A dynamic allocation index for the sequential design of experiments. In J. Gans (Ed.), *Progress in statistics* (pp. 241–266). Amsterdam, the Netherlands: North-Holland.

Herath, P., Klingberg, T., Young, J., Amunts, K., & Roland, P. (2001). Neural correlates of dual task interference can be dissociated from those of divided attention: An fMRI study. *Cerebral Cortex, 11,* 796–805. http://dx.doi.org/10.1093/cercor/11.9.796

Hills, T. T., Todd, P. M., & Goldstone, R. L. (2010). The central executive as a search process: Priming exploration and exploitation across domains. *Journal of Experimental Psychology: General, 139,* 590–609. http://dx.doi.org/10.1037/a0020666

Hommel, B. (2012). Convergent and divergent operations in cognitive search. In P. M. Todd, T. T. Hills, & T. W. Robbins (Eds.), *Cognitive search: Evolution, algorithms, and the brain. Strüngmann Forum Reports* (Vol. 9). Cambridge, MA: MIT Press.

Humphries, M. D., Khamassi, M., & Gurney, K. (2012). Dopaminergic control of the exploration-exploitation trade-off via the basal ganglia. *Frontiers in Neuroscience, 6,* 9. http://dx.doi.org/10.3389/fnins.2012.00009

Jepma, M., & Nieuwenhuis, S. (2011). Pupil diameter predicts changes in the exploration-exploitation trade-off: Evidence for the adaptive gain theory. *Journal of Cognitive Neuroscience, 23,* 1587–1596. http://dx.doi.org/10.1162/jocn.2010.21548

Kayser, A. S., Mitchell, J. M., Weinstein, D., & Frank, M. J. (2015). Dopamine, locus of control, and the exploration-exploitation tradeoff. *Neuropsychopharmacology, 40,* 454–462. http://dx.doi.org/10.1038/npp.2014.193

Khamassi, M., Enel, P., Dominey, P. F., & Procyk, E. (2013). Medial prefrontal cortex and the adaptive regulation of reinforcement learning parameters. *Progress in Brain Research, 202,* 441–464. http://dx.doi.org/10.1016/B978-0-444-62604-2.00022-8

Koechlin, E., Ody, C., & Kouneiher, F. (2003). The architecture of cognitive control in the human prefrontal cortex. *Science, 302,* 1181–1185. http://dx.doi.org/10.1126/science.1088545

Konstantinou, N., & Lavie, N. (2013). Dissociable roles of different types of working memory load in visual detection. *Journal of Experimental Psychology: Human Perception and Performance, 39,* 919–924. http://dx.doi.org/10.1037/a0033037

Krueger, P. M. (2017). Strategies for exploration in the domain of losses. *Judgment and Decision Making, 12,* 104–117.

Mars, R. B., Sallet, J., Rushworth, M. F., & Yeung, N. (Eds.). (2011). *Neural basis of motivational and cognitive control*. Cambridge, MA: MIT Press. http://dx.doi.org/10.7551/mitpress/9780262016438.001.0001

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., . . . Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision, 2,* 191–215. http://dx.doi.org/10.1037/dec0000033

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24,* 167–202. http://dx.doi.org/10.1146/annurev.neuro.24.1.167

Otto, A. R., Knox, W. B., Markman, A. B., & Love, B. C. (2014). Physiological and behavioral signatures of reflective exploratory choice. *Cognitive, Affective & Behavioral Neuroscience, 14,* 1167–1183. http://dx.doi.org/10.3758/s13415-014-0260-4

Payzan-LeNestour, E., & Bossaerts, P. (2011). Risk, unexpected uncertainty, and estimation uncertainty: Bayesian learning in unstable settings. *PLoS Computational Biology, 7,* e1001048. http://dx.doi.org/10.1371/journal.pcbi.1001048

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning: Current research and theory* (pp. 64–99). New York, NY: Appleton-Century-Crofts.

Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society, 58,* 527–536. http://dx.doi.org/10.1090/S0002-9904-1952-09620-8

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464. http://dx.doi.org/10.1214/aos/1176344136

Somerville, L. H., Sasse, S. F., Garrad, M. C., Drysdale, A. T., Abi Akar, N., Insel, C., & Wilson, R. C. (2017). Charting the expansion of strategic exploratory behavior during adolescence. *Journal of Experimental Psychology: General, 146,* 155–164. http://dx.doi.org/10.1037/xge0000250

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Warren, C. M., Wilson, R. C., van der Wee, N. J., Giltay, E. J., van Noorden, M. S., Cohen, J. D., & Nieuwenhuis, S. (2017). The effect of atomoxetine on random and directed exploration in humans. *PLoS ONE, 12,* e0176034. http://dx.doi.org/10.1371/journal.pone.0176034

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans use directed and random exploration to solve the explore-exploit dilemma. *Journal of Experimental Psychology: General, 143,* 2074–2081. http://dx.doi.org/10.1037/a0038199

Wilson, R. C., & Niv, Y. (2012). Inferring relevance in a changing world. *Frontiers in Human Neuroscience, 5,* 189. http://dx.doi.org/10.3389/fnhum.2011.00189

Yu, A. J., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron, 46,* 681–692. http://dx.doi.org/10.1016/j.neuron.2005.04.026

Zajkowski, W. K., Kossut, M., & Wilson, R. C. (2017). A causal role for right frontopolar cortex in directed, but not random, exploration. *eLife, 6,* e27430. http://dx.doi.org/10.7554/eLife.27430

Received January 22, 2018
Revision received October 26, 2018
Accepted October 31, 2018 ■