

1 Choice seeking is motivated by the intrinsic need for personal control

2 **Authors**

3 *Jérôme Munuera^{1,2,*}, Marta Ribes Agost², David Bendetowicz¹, Adrien Kerebel², Valérian*
4 *Chambon^{2,†,*}, Brian Lau^{1,†,*}*

5 **Affiliations**

6 *1. Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, Inserm, CNRS, APHP,*
7 *Paris, France*

8 *2. Institut Jean Nicod, Département d'études cognitives, ENS, EHESS, CNRS, PSL University,*
9 *75005 Paris, France*

10 *†. Co-last authors*

11 **. Corresponding authors: jerome.munuera@icm-institute.org; valerian.chambon@ens.fr;*
12 *brian.lau@upmc.fr*

13 **Keywords:** decision-making; reward; reinforcement learning; human; agency

14 **Abstract:**

15 When deciding between options that do or do not lead to future choices, humans often choose to
16 choose. We studied choice seeking by asking subjects to decide between a choice opportunity or
17 performing a computer-selected action. Subjects preferred choice when these options were equally
18 rewarded, even deterministically, and were willing to trade extrinsic rewards for the opportunity to
19 choose. We explained individual variability in choice seeking using reinforcement learning models
20 incorporating risk sensitivity and overvaluation of rewards obtained through choice. Degrading
21 perceived controllability diminished choice preference, although willingness to repeat selection of
22 choice opportunities remained unchanged. In choices following these repeats, subjects were
23 sensitive to rewards following freely chosen actions, but ignored environmental information in a
24 manner consistent with a desire to maintain personal control. Choice seeking appears to reflect the
25 intrinsic need for personal control, which competes with extrinsic reward properties and external
26 information to motivate behavior.

27 **Author summary:**

28 Human decisions can often be explained by the balancing of potential rewards and punishments.
29 However, some research suggests that humans also prefer opportunities to choose, even when
30 these have no impact on future rewards or punishments. Thus, opportunities to choose may be
31 intrinsically motivating, although this has never been experimentally tested against alternative
32 explanations such as cognitive dissonance or exploration. We conducted behavioral experiments
33 and used computational modelling to provide compelling evidence that choice opportunities are
34 indeed intrinsically rewarding. Moreover, we found that human choice preference varied
35 according to individual risk attitudes, and expressed a need for personal control that competes
36 with maximizing reward intake.

45 Preference for choice has been observed in humans(1–6) as well as other animals including rats(7),
46 pigeons(8) and monkeys(9,10). This free-choice premium can be behaviorally measured by having
47 subjects perform trials in two stages: a decision is first made between the opportunity to choose
48 from two terminal actions (*free*) or to perform a mandatory terminal action (*forced*) in the second
49 stage(7). Although food or fluid rewards follow terminal actions in non-human studies, choice
50 preference in humans can be elicited using hypothetical outcomes that are never obtained(3,11).
51 Thus, choice opportunities appear to possess or acquire value in and of themselves. It may be that
52 choice has value because it represents an opportunity to exercise control, which is itself intrinsically
53 rewarding(1,4,12). Personal control is central in numerous psychological theories, where
54 constructs such as autonomy(13,14), controllability(15,16), personal causation(17), effectance(18),
55 perceived behavioral control(19) or self-efficacy(15) are key for motivating behaviors that are not
56 economically rational or easily explained as satisfying basic drives such as hunger, thirst, sex, or
57 pain avoidance(20).

58 There are alternative explanations for choice seeking. For example, subjects may prefer
59 choice because they are curious and seek information(21,22), or they wish to explore potential
60 outcomes to eventually exploit their options(23), or because they seek variety to perhaps reduce
61 boredom(24) or keep their options open(3). By these accounts, however, the expression of personal
62 control is not itself the ends, but rather a means for achieving an objective that once satisfied
63 reduces choice preference. For example, choice preference should decline when there is no further
64 information to discover in the environment, or after uncertainty about reward contingencies have
65 been satisfactorily resolved.

66 Choice seeking may also arise due to selection itself altering outcome representations.
67 Contexts signaling choice opportunities may acquire distorted value through choice-induced
68 preference change(25). By this account, deciding between equally valued terminal actions

69 generates cognitive dissonance that is resolved by post-choice revaluation favoring the chosen
70 action(25,26). This renders the free option more valuable than the forced option since revaluation
71 only occurs for self-determined actions(27,28). Alternatively, subjects may develop distorted
72 outcome representations through a process related to the winner's or optimizer's curse(29),
73 whereby optimization-based selection upwardly biases value estimates for the chosen action. One
74 algorithm subject to this bias is Q-learning(30), where action values are updated using the
75 maximum value to approximate the maximum expected value. In a two-stage task, the free action
76 value is biased upwards due to considering only the best of two possible future actions, while the
77 forced action value remains unbiased since there is only one possible outcome(31). Again, the
78 expression of personal control is not itself the ends for these selection-based accounts, and both
79 predict that choice preference should be reduced when terminal rewards associated with the free
80 option are clearly different.

81 Data from prior studies does not arbitrate between competing explanations for choice-
82 seeking. Here, we used behavioral manipulations and computational modelling to explore the
83 factors governing human preference for choice. In the first experiment, we altered the reward
84 contingencies associated with terminal actions in order to rule out curiosity, exploration, variety-
85 seeking, and selection-based explanations for choice seeking. In the second experiment, we used a
86 titration procedure to measure the value of choice relative to an extrinsic reward available in the
87 environment (i.e., money). We then used reinforcement learning models to show that optimistic
88 learning (considering the best possible future outcome) was insufficient to explain individual
89 variability in choice seeking. Rather, subjects adopted different *decision attitudes*, the desire to
90 make or avoid decisions independent of the outcomes(11), which were balanced against differing
91 levels of risk sensitivity. Finally, in the third experiment, we sought to test whether choice
92 preference was motivated by personal control beliefs. We manipulated the perceived controllability

93 of the task and found that subjects' willingness to repeat a free choice was not affected by the lack
94 of objective controllability over reward outcome. Importantly, subjects were sensitive to past
95 rewards only in trials where state outcomes could be attributed to self-determined choice, and
96 ignored rewards on trials where there was an apparent loss of control. Together, our results support
97 the hypothesis that human preference for choice opportunities derives from the intrinsic motivation
98 for personal control.

99 **Results:**

100 Subjects performed repeated trials with a two-stage structure (Fig. 1). In each trial, subjects made
101 a 1st-stage choice between two options defining the 2nd-stage: the opportunity to choose between
102 two fractal targets (*free*) or performing an obligatory selection of another fractal target (*forced*).
103 Extrinsic rewards (€) were delivered only for terminal (i.e., 2nd-stage) actions. If subjects chose the
104 *forced* option, the computer always selected the same fractal target for the subjects. If subjects
105 chose the *free* option, they had to choose between two fractal targets associated with two different
106 terminal states. We fixed reward contingencies in blocks of trials, and used unique fractal targets
107 for each block. We divided each block into an initial training phase (Fig. 1B) followed by a test
108 phase (Fig. 1C) to ensure that the subjects learned the associations between the different fractal
109 targets and extrinsic reward probabilities.

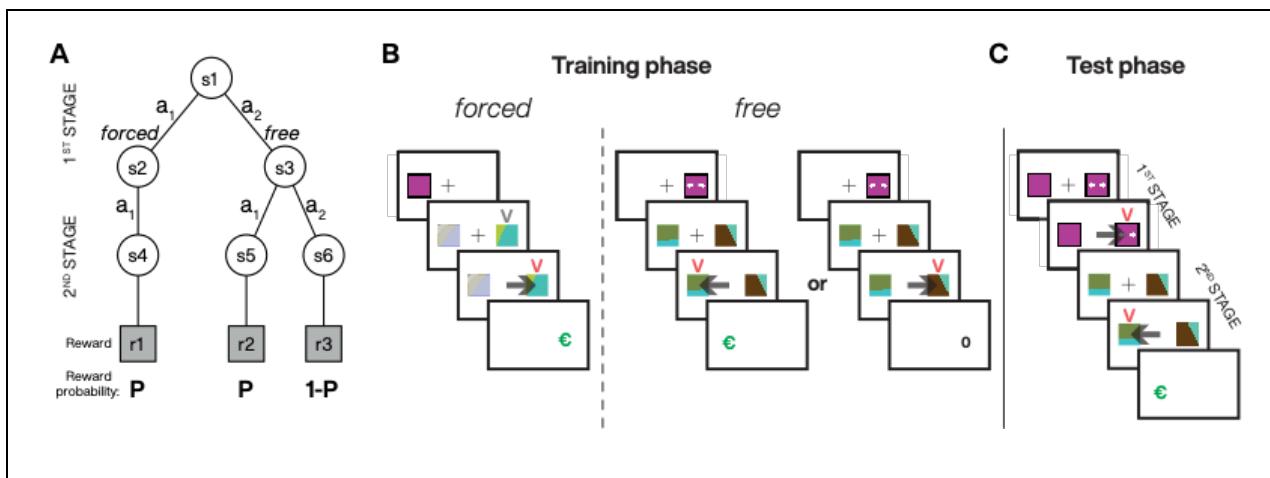


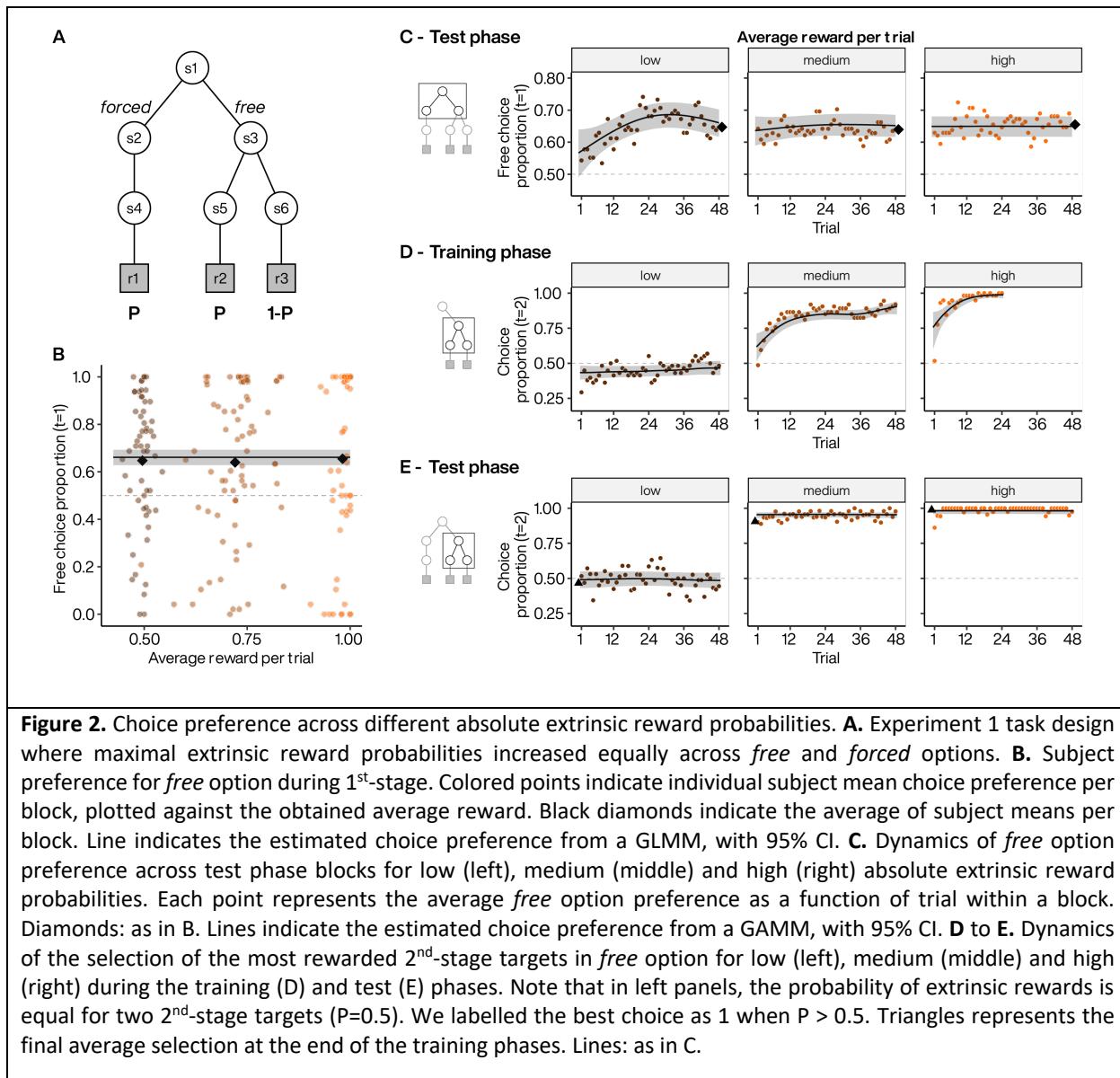
Figure 1. Two-stage task structure. **A.** State diagram illustrating the 6 possible states (s), actions (a) and associated extrinsic reward probabilities (e.g., $P = 0.5, 0.75$ or 1 for blocks 1 to 3, respectively); s_2 and s_3 were represented by two different 1st-stage targets (e.g., colored squares with or without arrows for *free* and *forced* trials, respectively) and s_4 to s_6 were associated to three different 2nd-stage targets (fractals). **B.** Sequence of events during the training phase where the subjects learned the contingencies between the fractal targets and their reward probabilities (P) associated with the *forced* (no choice) and *free* (choice available) options. When training the reward contingencies associated with the *forced* option, subjects' actions in the 2nd-stage had to match the target indicated by a grey V-shape and was always the same (s_4). When training the reward contingencies associated with the *free* option, no mandatory target is present at the 2nd-stage (s_5 or s_6 can be chosen) but one of the targets is more rewarded when $P > 0.5$. Black arrows represent the selection of the target by the subject. **C.** Sequence of events during the test phase: subjects first decided between the *free* or *forced* option and then experienced the associated 2nd-stage. Rewards, when delivered, were represented by a large green euro symbol (€).

110 **Free choice preference across different extrinsic reward probabilities**

111 In experiment 1, we varied the overall expected value by varying the probability of extrinsic reward
112 delivery (P) across different blocks of trials. These probabilities ranged from 0.5 to 1 across the
113 blocks (i.e., low to high), and the programmed probabilities in *free* and *forced* 2nd-stage rewards
114 were equal (Fig. 2A). For example, in high probability blocks, we set the probabilities of the *forced*
115 terminal action and of one of the *free* terminal actions (a1) to 1, and set the probability of the second
116 *free* terminal action (a2) to 0. Therefore, the maximum expected value was equal for the *free* and
117 *forced* options.

118 Subjects chose to choose more frequently, selecting the *free* option in 64% (n=58) of test
119 trials on average (Fig. 2B). The level of preference did not differ significantly across blocks ($p =$
120 0.857, low = 65%, medium = 64%, high = 66%). We found that subjects immediately expressed
121 above chance preference for the *free* option (Fig. 2C) despite never having actualized 1st-stage
122 choices during training. Looking within a block, we found that subjects' preference remained
123 constant across trials in medium and high reward probability blocks ($p = 0.22$ and 0.6823 for
124 nonlinear smooth by trial deviating from a flat line, respectively; Fig. 2C, middle and right panels).
125 In low probability blocks, subjects started with a lower choice preference that gradually increased
126 to match that observed in the medium and high probability blocks ($p = 0.0014$ for nonlinear smooth
127 by trial; Fig. 2C left panel). The lower reward probability may have prevented subjects from
128 developing accurate reward representations by the end of the training phase, which may have led
129 to additional sampling of the three 2nd-stage targets (two in *free* and one in *forced*) in the beginning
130 of the test phase.

131



132

133 **Second-stage performance following *free* selection**

134 We investigated participants' 2nd-stage choices following *free* selection to exclude the possibility
 135 that choice preference arose because reward contingencies had not been learned. During the
 136 training phase, when $P > 0.5$, participants quickly learned to choose the most rewarded fractal targets
 137 (at $P=0.5$, all fractal targets were equally rewarded) (Fig. 2D). During the test phase, participants

138 continued to select the same targets (Fig. 2E), confirming stable application of learned
139 contingencies ($p > 0.1$ for nonlinear smooth by trial deviating from a flat line for all blocks).

140 Choice preference was not explained by subjects obtaining more extrinsic rewards
141 following selection of *free* compared to *forced* options. Obtained reward proportions were not
142 significantly different in the low (following selection of *free* vs. *forced*, 0.516 vs. 0.536, $p = 0.276$)
143 or medium (0.746 vs. 0.762, $p = 0.322$) probability blocks. In contrast, in high probability blocks,
144 subjects received significantly fewer rewards on average after *free* selection than after *forced*
145 selection (0.989 vs. 1, $p = 0.0016$). In this block, reward was fully deterministic, and *forced*
146 selection always led to a reward, whereas *free* selections could lead to missed rewards if subjects
147 chose the incorrect target.

148

149 **Trading extrinsic rewards for choice opportunities**

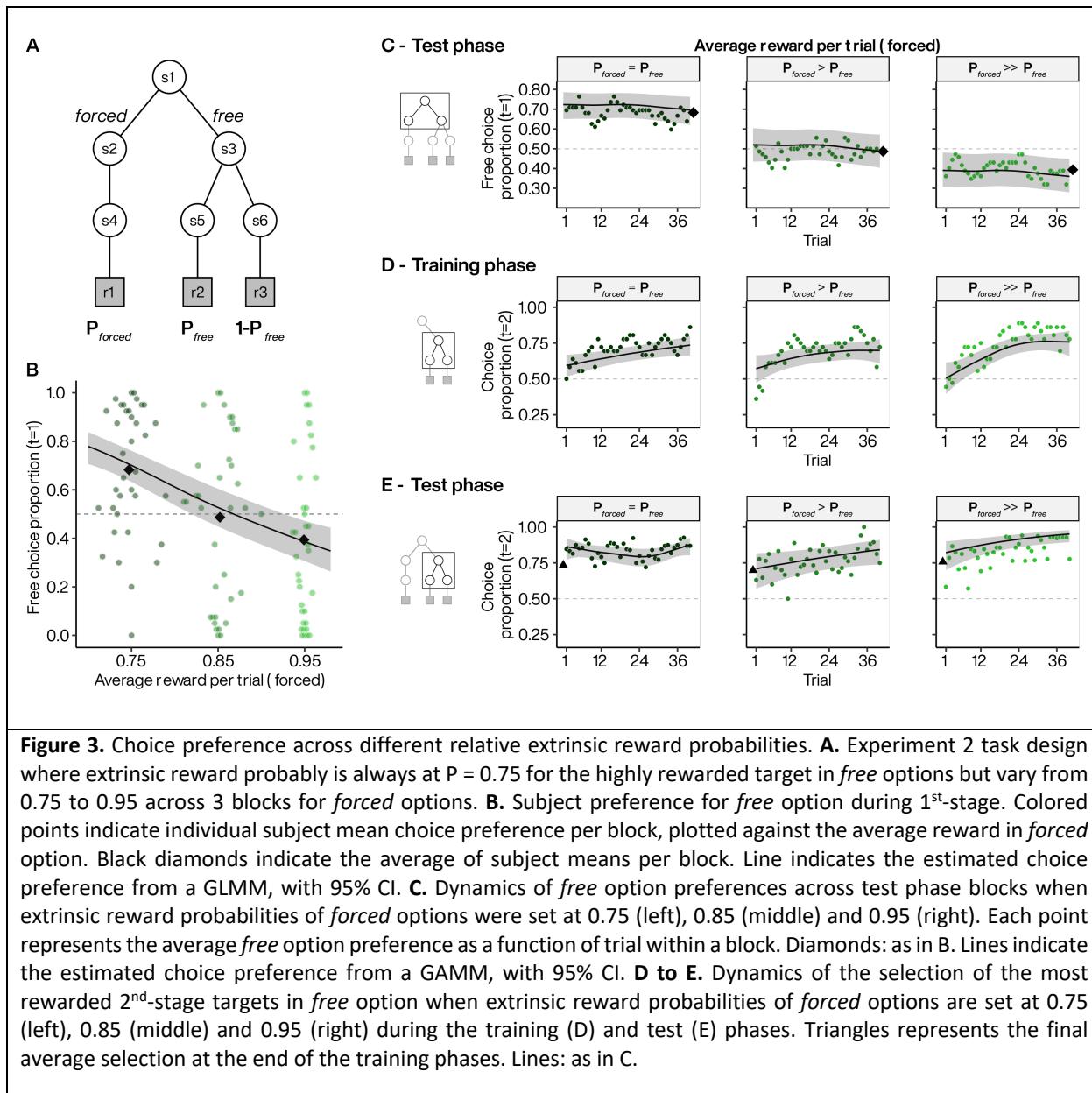
150 Since manipulating the overall expected reward did not alter choice seeking behavior at the group-
151 level, we investigated the effect of changing the relative expected reward between 1st-stage options.
152 In experiment 2, we tested a new group of 36 subjects for whom we decreased the objective value
153 of the *free* versus *forced* options. This allowed us to assess the point at which these options were
154 equally valued and potentially reversed to favor the initially non-preferred (*forced*) option (Fig.
155 3A). Thus, we titrated the value of choice opportunity by increasing the reward probabilities
156 following *forced* selection (block 1: $P_{forced} = 0.75$; block 2: $P_{forced} = 0.85$; block 3: $P_{forced} = 0.95$),
157 while keeping the reward probabilities following *free* selection fixed ($P_{free}|a1 = 0.75$, $P_{free}|a2 = 0.25$
158 for all blocks).

159 As in experiment 1, we found that subjects preferred choice when the extrinsic reward
160 probabilities of the *free* and *forced* options were equal (block 1: 68% 1st-stage choice in favor of
161 *free*; Fig. 3B, dark green). Increasing the reward probability associated with the *forced* option

162 significantly reduced choice preference ($p = 0.00344$, Fig. 3B) to 49% (block 2) and 39% (block
163 3). We estimated the population preference reversal point at $P_{forced} = 0.88$, indicating that
164 indifference was obtained on average when the value of the *forced* option was 17% greater than
165 that of the *free*. We found that subjects' preference remained constant across trials when reward
166 probabilities were equal ($p = 0.875$ for nonlinear smooth by trial; Fig. 3C, left panel). Although
167 reduced overall, the selection of the *free* option also did not vary across trials in blocks 2 and 3 (p
168 = 0.737 and 0.078 for nonlinear smooth by trial, respectively). Furthermore, as in experiment 1,
169 subjects acquired preference for the most rewarded 2nd-stage targets during the learning phase
170 (Fig. 3D) and continued to express this preference during the test phase in all three blocks (Fig. 3E).
171 Thus, the decrease in choice preference was not related to failure to learn the reward contingencies
172 during the training phase.

173 Although decreasing the relative value of the *free* option reduced choice preference, most
174 subjects did not switch exclusively to the *forced* option. Even in block 3, where the *forced* option
175 was set to be rewarded most frequently ($P_{forced} = 0.95$ versus $P_{free} = 0.75$), 32/36 subjects selected
176 the *free* option in a non-zero proportion of trials. Since exclusive selection of the *forced* option
177 would maximize extrinsic reward intake, continued *free* selection indicates a persistent appetency
178 for choice opportunities despite their diminished relative extrinsic value.

179



180

181 **Reinforcement-learning model of choice seeking**

182 We next sought to explain individual variability in choice behavior using a value-based decision-
 183 making framework. We first used mixed logistic regression to examine whether rewards obtained
 184 from 2nd-stage actions influenced 1st-stage choices. We found that obtaining a reward on the
 185 previous trial significantly increased the odds that subjects repeated the 1st-stage selection that
 186 ultimately led to that reward ($p < 0.0001$, odds ratio rewarded/unrewarded on previous trial: 1.92

187 $\pm 95\%$ CI [1.40, 2.60]). This suggest that subjects continued to update their extrinsic reward
188 expectation based on experience during the test phase. We therefore leveraged the framework of
189 temporal-difference reinforcement learning (TDRL) to provide a model-based characterization of
190 the emergence of choice preference.

191 We fitted TDRL models to individual data using two distinct features to capture individual
192 variability across different extrinsic reward contingencies. The first feature was a free choice bonus
193 added to self-determined actions as an intrinsic reward. This can lead to overvaluation of the *free*
194 option via standard TD learning. The second feature modifies the form of the future value estimate
195 used in the TD value iteration, which in common TDRL variants is, or approximates, the best future
196 action value (Q-learning or SARSA with softmax behavioral policy, respectively). We treated both
197 Q-learning and SARSA together as optimistic algorithms since they are not highly discriminable
198 with our data (Supplementary Fig. 1). We compared this optimism with another TDRL variant that
199 explicitly weights the best and worst future action values (Gaskett's β -pessimistic model(32)),
200 which could capture avoidance of choice opportunities through increased weighting of the worst
201 possible future outcome (pessimistic risk attitude). For example, risk is maximal in the high reward
202 probability block in experiment 1 since selection of one 2nd-stage target led to a guaranteed reward
203 (best possible outcome) whereas selection of the other target led to guaranteed non-reward (worst
204 possible outcome).

205 We found that it was necessary to incorporate the overvaluation of rewards obtained from
206 *free* actions to predict choice preference in experiment 1 (Fig. 4A). Moreover, the magnitude of
207 the bonus was significantly associated with increasing choice preference during the 1st-stage of the
208 trials ($p = 0.0005$ for nonlinear smooth, Fig. 4B). Therefore, optimistic or pessimistic targets alone
209 were insufficient to explain individual choice preference across different extrinsic reward
210 contingencies. We found that a pessimistic target best fitted about 28% (16 of 58) of the subjects

211 in experiment 1. Moreover, most pessimistic subjects (13 of 16) were best fitted with a model
 212 including a free choice bonus to balance risk and decision attitudes across reward contingencies.
 213 In experiment 1, we introduced risk by varying the difference in extrinsic reward probability for
 214 the best and worst outcome following *free* selection. The majority of so-called ‘pessimistic
 215 subjects’ preferred choice when extrinsic reward probabilities were low, but their weighting of the
 216 worst possible outcome decreased this preference as risk increased (Fig. 4C, pink). Thus,
 217 pessimistic subjects avoided the *free* option despite rarely or never selecting the more poorly
 218 rewarded 2nd-stage target during the test phase.

219 We also fitted the TDRL variants to individual data from experiment 2, and found that a
 220 free choice bonus was also necessary to explain choice preference across extrinsic reward
 221 contingencies in that experiment. Four subjects (of 36) were best fitted using the β -pessimistic
 222 target (see Supplementary Fig. 2) although this may be a conservative estimate since we did not
 223 vary risk in experiment 2.

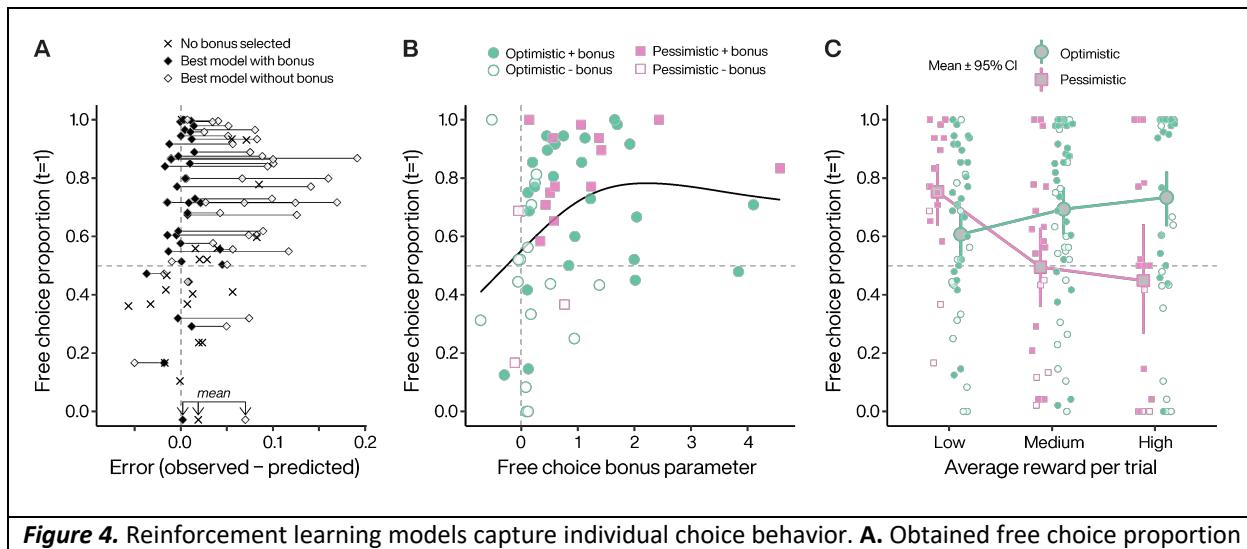


Figure 4. Reinforcement learning models capture individual choice behavior. **A.** Obtained free choice proportion as a function of model error in experiment 1, averaged over all conditions. For subjects where the selected model did not include a free choice bonus, only one symbol (X) is plotted. For subjects where the selected model included a free choice bonus, two symbols are plotted. Filled symbol represents the fit error with the selected model, and the open symbol represents the next best model that did not include a free choice bonus. Lines connect individual subjects. **B.** Bonus coefficients increase as a function of subjects’ preference for *free* options irrespectively of the target policy they used when performing the task. Choice preference from low probability blocks ($P=0.5$). Filled circles indicate that the best model included a free choice bonus parameter. Line illustrates a generalized additive

model smooth. **C.** Pessimistic subjects significantly decrease their *free* option preference as a function of extrinsic reward probabilities. Symbol legend from B applies to the small points representing individual means in C. Error bars for 95% CI.

224

225 **Influence of action-outcome coherence on choice seeking behavior**

226 We next asked whether choice preference was related to personal control beliefs. To do so, we
227 manipulated the coherence between an action and its consequence over the environment. In
228 experiment 3, we tested the relationship between preference for choice opportunity and the physical
229 coherence of the terminal action by directly manipulating the perceived controllability of 2nd-stage
230 actions. We modified the two-stage task to introduce a mismatch between the subject's selection
231 of the 2nd-stage target and the target ultimately displayed on the screen by the computer (Fig. 5A).
232 We did this by manipulating the probability that a 2nd-stage target selected by a subject would be
233 swapped for the 2nd-stage target that had not been selected. That is, on coherent trials, a subject
234 selecting the fractal on the right side of the screen would receive visual feedback indicating that
235 the right target had been selected. On incoherent trials, a subject selecting the fractal on the right
236 side would receive feedback that the opposite fractal target had been selected (i.e., the left target).

237 To ensure that all other factors were equalized between the two 1st-stage choices, we
238 implemented target swaps following both *free* and *forced* selections by adding an additional state
239 to our task (Fig. 5A). In one block of trials, the incoherence was set to 0 and every subject action
240 in the 2nd-stage led to a coherent selection of the second target. In the other blocks, we set
241 incoherence to 0.15 or 0.3, resulting in lower perceived controllability between target choice and
242 target selection (e.g., 85% of the time, pressing the left key will select the left target, and in 15%
243 the right target). We set all of the extrinsic reward probabilities associated with the different fractal
244 targets to $P = 0.75$. Since all 2nd -stage actions had the same expected value, the experiment was
245 objectively uncontrollable because the probability of reward was independent of all actions(16).

246 Moreover, equal reward probabilities ensured that outcome diversity(33,34), outcome entropy(35),
247 and instrumental divergence(36) did not contribute to choice preference since these were all equal
248 between the *forced* and *free* options.

249 The same group of participants who performed experiment 2 also performed experiment 3
250 (n=36). Choice preference was high (70%) in block 1 when coherence was not altered, similar to
251 block 1 from experiment 2 where extrinsic reward was equal between *free* and *forced* options. The
252 only difference between these two blocks was that choosing the *forced* option resulted in the
253 obligatory selection of the same fractal (experiment 2) or one of two fractals randomly selected by
254 the computer (experiment 3), which indicates that subjects' choice preference was not related to
255 action variability per se following *forced* selection. Moreover, we found that choice preference was
256 significantly correlated ($r = 0.358$, $p = 0.03175$) between block 1 of experiments 2 and 3,
257 highlighting a within-subject consistency in choice preference.

258 Increasing the incoherence of the 2nd-stage actions progressively reduced choice preference
259 (block 2 and 3: 67% and 64% in favor of *free* respectively). As in experiments 1 and 2, choice
260 preference was expressed immediately after the training phase and remained constant throughout
261 the different blocks (Supplementary Fig. 3). We found that the decline in choice preference
262 depended on the 1st-stage choice on the previous trial. Specifically, following coherent trials, we
263 found that there was a significant interaction between the previous 1st-stage choice (*free* or *forced*)
264 and the degree of incoherence ($p = 0.0015$, Fig. 5B). The difference in slopes was due to decreasing
265 propensity to choose the *free* option following *forced* selection on the previous trial ($p = 0.0111$),
266 with no change in the propensity to choose the *free* option following *free* selection on the previous
267 trial ($p = 0.8706$). Thus, as incoherence increased, subjects tended to stay more with the *forced*
268 option, while maintaining a preference to repeat *free* selections.

269 The sustained repetition of *free* selections across the different levels of incoherence
270 suggests that subjects may have been seeking to regain control of the environment through self-
271 determined 2nd-stage choices. Although the task was objectively uncontrollable since all terminal
272 action-target sequences were associated with the same reward probability, subjects may have
273 developed structure beliefs based on local reward history and target swaps, which could be reflected
274 in 2nd-stage patterns of choice. Thus, subjects may have followed a strategy based on reward
275 feedback by repeating only actions associated with a previous reward (illusory maximization of
276 reward intake; Fig.5C, first panel). Alternatively, they could have followed a strategy based on
277 action-outcome incoherence feedback and thus avoided trials associated with a previous target
278 swap (illusory minimization of incoherent states; Fig. 5C, second panel). However, subjects may
279 have also employed another classic strategy known as “model-based” where agents use their (here
280 illusory) understanding of the task structure built from all the information provided by the
281 environment (Fig.5C, third panel)(37). Under this strategy, subjects try to integrate both the reward
282 and target-swap feedback to select the next target in order to maximize reward. For example, an
283 incoherent but rewarded trial would lead to a behavioral switch because the subject has integrated
284 the information provided by the environment (i.e., the target swap induced by the computer),
285 signaling that the other target is actually rewarded (see second bar on third panel of Fig. 5C).
286 Finally, an alternative strategy could rely on maximizing personal (i.e., internal) control, where the
287 subject is the (illusory) agent of the entire sequence of events (i.e., action-state-reward) and would
288 therefore ignore reward outcomes when they are not associated with the selected action-state
289 (Fig.5C, fourth panel).

290 Results of the stay behavior during 2nd-stage choice following *free* selection suggests that
291 subjects seek personal control when choosing between the different fractal targets (Fig.5D). Indeed,
292 when their action was consistent with the state they were choosing (i.e., the coherent fractal target

293 feedback), they took the reward outcome into account to adjust their behavior on the next trial,
 294 either by staying on the same target when the trial was rewarded or by switching to the other one
 295 when no reward was delivered. However, subjects were insensitive to the reward outcome during
 296 incoherent trials as they maintained the same strategy (staying) during subsequent trials, regardless
 297 of whether they were previously rewarded or not. This strategy reflects an attempt to regain
 298 personal control over the environment at the expense of the task goal of maximizing reward intake.
 299

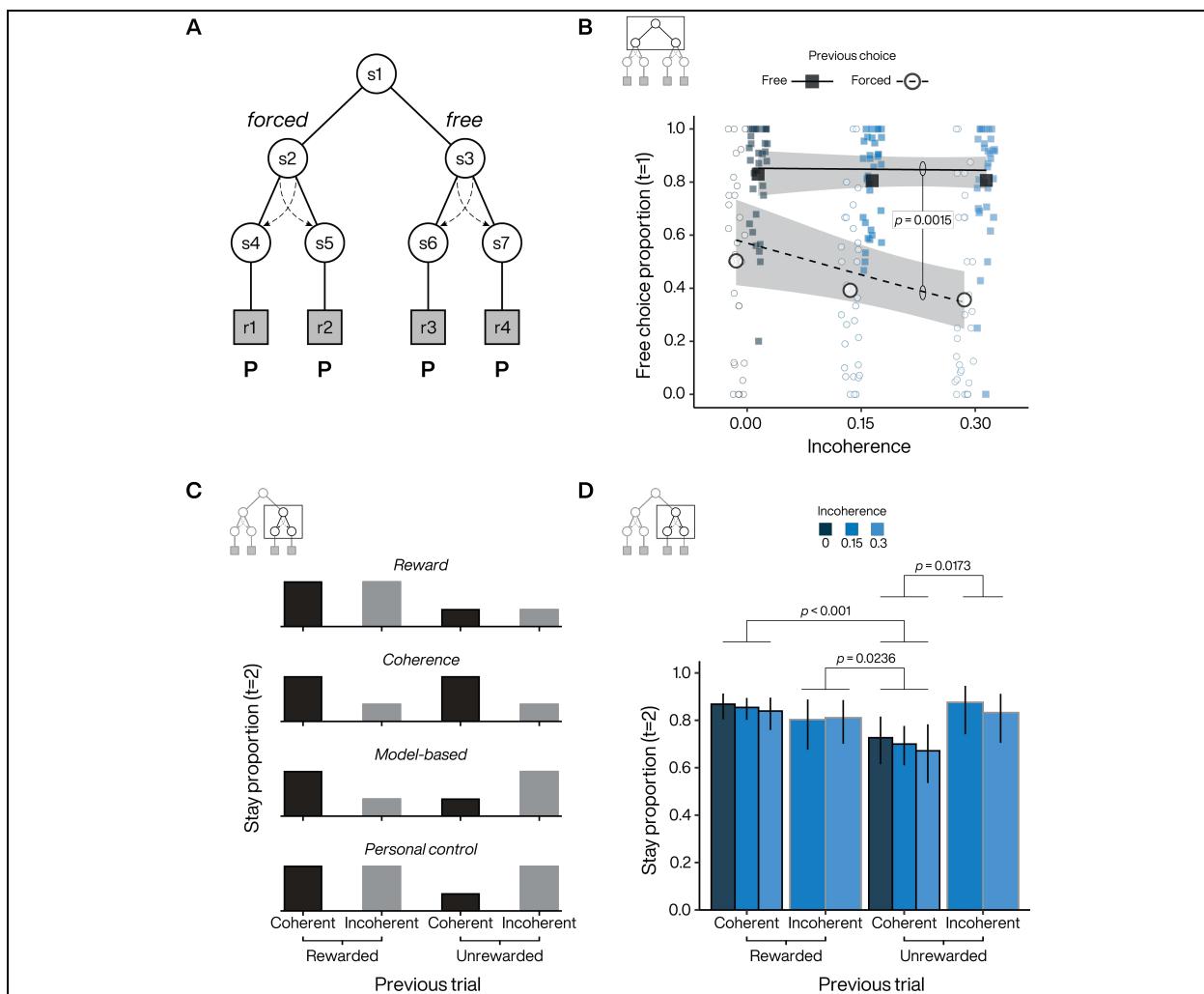


Figure 5. Perceived controllability alters choice preference. **A.** Task design where a 7th state, associated to the *forced* options, has been added to manipulate the incoherence in both *free* and *forced* options. At incoherence = 0, the visual feedback presented to the subject matches their selected target. Extrinsic reward probabilities set at $P=0.75$ for all the 2nd-stage targets. **B.** First-stage probabilities to stay or switch in free options after a *free* and *forced* trial respectively, as a function of the different incoherence blocks. **C.** Second-stage stay probabilities for

the different action-state-reward trial type. Each sub-panels represent a putative strategy followed by the subject.

D. Estimated 2nd-stage stay probabilities. Error bars for 95% CI. P-values are displayed for significant pairwise comparisons and adjusted for multiple comparisons.

300

301

302 **Discussion**

303 Animals prefer situations that offer more choice to those that offer less. Although this behavior can
304 be reliably measured using the two-stage task design popularized by Voss and Homzie(7), their
305 conclusion that choice has intrinsic value is open to debate. To rule out alternative explanations for
306 choice-seeking, we performed three experiments in which we clearly separated learning of reward
307 contingencies from testing of choice preference. Our experiments point to a sustained preference
308 for choice opportunities that express an intrinsic need for personal control. Moreover, this need
309 may compete with potentially valuable information for maximizing outcomes or even extrinsic
310 rewards per se.

311 In the first and second experiments, we varied the reward probabilities associated with
312 terminal actions following *free* and *forced* selection. Consistent with previous studies, subjects
313 preferred the opportunity to make a choice when expected rewards were equal between terminal
314 actions ($P = 0.5$). Surprisingly, subjects also preferred choice when we increased the value
315 difference between terminal actions in the *free* option, while keeping the *maximum* expected reward
316 equal in the free and forced options ($P > 0.5$). This sustained preference for choice is potentially
317 economically suboptimal since making a free choice carries the risk of making an error leading to
318 lowered reward intake. The persistence of this preference for free choice even when reward
319 delivery was deterministic ($P = 1$), makes it unlikely that this preference was due to an
320 underestimation of forced trials due to poor learning of reward contingencies.

321 Subjects appeared to have understood the reward contingencies as evidenced by their
322 consistent preference for the highest-rewarded 2nd-stage fractal, which was acquired during the
323 training phase and expressed during the test phase. This stable 2nd-stage fractal selection, together
324 with the immediate expression and maintenance of 1st-stage choice preference, renders unlikely

325 accounts based on curiosity, exploration or variety seeking since varying the probability of rewards
326 did not modulate choice preference about two third of the subjects (i.e., optimistic subjects).

327 Selection-based accounts also have trouble explaining the pattern of results we observed.
328 The idea that post-choice revaluation specifically inflates expected outcomes after choosing the
329 free option can explain choice-seeking when all terminal reward probabilities are equal. However,
330 post-choice revaluation cannot explain choice preference when the terminal reward probabilities
331 in the *free* option clearly differ from one another, since revaluation appears to occur only after
332 choosing between closely valued options(28,38). That is, there is no cognitive dissonance to resolve
333 when reward contingencies are easy to discriminate, and no preference for choice should be
334 observed when the choice is between a surely (i.e., deterministically) rewarded action and a never
335 rewarded action. The existence of choice preference in the deterministic condition ($P = 1$) also
336 cannot be explained by an optimistic algorithm such as Q-learning, since the maximum action value
337 is equal to the maximum expected value, and the value of the free option is not biased upwards
338 under repeated sampling(31).

339 Although standard Q-learning could not capture variability across different terminal reward
340 probabilities, we found that combining two novel modifications to TDRL models was able to do
341 so. The first feature was a free choice bonus—a fixed value added to all extrinsic rewards obtained
342 through free actions—that can lead to overvaluation of the free option via standard TD learning.
343 This bonus implements Beattie and colleagues' concept of *decision attitude*, the desire to make or
344 avoid decisions independent of the outcomes(11). The second feature modifies the form of the
345 future value estimate in the TD value iteration. Zorowitz and colleagues(31) showed that replacing
346 the future value estimate in Q-learning with a weighted mixture of the best and worst future action
347 values(32) can generate behavior ranging from aversion to preference for choice. The mixing
348 coefficient determines how optimism (maximum of future action values, total risk indifference) is

349 tempered by pessimism (minimum of future action values, total risk aversion). In experiment 1, we
350 found that 28% of subjects were best fitted with a model incorporating pessimism, which captured
351 a downturn in choice preference with increasing relative value difference between the terminal
352 actions in the *free* option. Importantly however, individual variability in the TD future value
353 estimates alone did not explain the pattern of choice preference across target reward probabilities,
354 and a free choice bonus was still necessary for most subjects. Thus, the combination of both a free
355 choice bonus (decision attitude) and pessimism (risk attitude) was key for explaining why some
356 individuals shift from seeking to avoiding choice. This was unexpected because the average choice
357 preference in experiment 1 was not significantly different across reward manipulations,
358 highlighting the importance of examining behavior at the individual level. Here, we examined risk
359 using the difference between the best and worst outcomes as well as relative value using probability
360 (see(39)). It may be the case that variability is also observed in how individuals balance the intrinsic
361 rewards with other extrinsic reward properties that can influence choice preference, such as reward
362 magnitude(39).

363 Why are choice opportunities highly valued? It may be that choice opportunities have
364 acquired intrinsic value because they are particularly advantageous in the context of the natural
365 environment in which the learning system has evolved. Thus, choice opportunities might be
366 intrinsically rewarding because they promote the search for states that minimize uncertainty and
367 variability, which could be used by an agent to improve their control over the environment and
368 increase extrinsic reward intake in the long run(40,41). Developments in reinforcement learning
369 and robotics support the idea that both extrinsic and intrinsic rewards are important for maximizing
370 an agent's survival(42–44). Building intrinsic motivation into RL agents can promote the search
371 for uncertain states and facilitate the acquisition of skills that generalize better across different
372 environments, an essential feature for maximizing an agent's ability to survive and reproduce over

373 its lifetime, i.e. its evolutionary fitness(42).

374 The intrinsic reward of choice may be a specific instance of more general motivational
375 constructs such as autonomy(13,14), personal causation(17), effectance(18), learned
376 helplessness(45), perceived behavioral control(19) or self-efficacy(15), which are key for
377 motivating behaviors that are not easily explained as satisfying basic drives such as hunger, thirst,
378 sex, or pain avoidance(20). Common across these theoretical constructs is that control is
379 intrinsically motivating only when the potential exists for agents to determine their own behavior,
380 which when realized can give rise to a sense of agency and, in turn, strengthens the belief in the
381 ability to exercise control over one's life(46). Thus, individuals with an *internal* locus of control
382 tend to believe that they, as opposed to external factors such as chance or other agents, control the
383 events that affect their lives. Crucially, the notion of locus of control makes specific predictions
384 about the relationship between preference for choice—choice being an opportunity to exercise
385 control—and the environment: individuals should express a weaker preference for choice when the
386 environment is adverse, stressful or unpredictable(47). This prediction is consistent with what is
387 known about the influence of environmental adversity on control externalization: individuals
388 exposed to greater environmental instabilities tend to believe that external and uncontrollable
389 forces are the primary causes of events that affect their lives, as opposed to themselves(48). In other
390 words, one would expect belief in one's ability to control events, and thus preference for choice, to
391 decline as the environment is perceived as increasingly unpredictable.

392 In our third experiment, we sought to test whether it was specifically a belief in personal
393 control that motivated subjects, by altering the perceived controllability of the task environment.
394 To do so, we introduced a novel change to the two-stage task where in a fraction of trials subjects
395 experienced random swapping of the terminal states (fractals). Thus, subjects were subjected to
396 trials where the terminal state was incoherent with their choice, and thus experienced alterations in

397 their ability to predict the state of the environment following their action. Incoherence occurred
398 with equal probability following free and forced actions in order to equate for any value associated
399 with swapping itself. We found a significant reduction in the propensity to switch from forced to
400 free choice following action-target incoherence, suggesting that altering the perceived
401 controllability of the task causes choice to lose its attractiveness. This reduction in choice
402 preference following incoherent trials is reminiscent of a form of locus externalization, and is
403 consistent with the notion that choice preference is driven by a belief in one's personal control. In
404 this experiment, we focused on the value of personal control, and therefore held other decision
405 variables such as outcome diversity(33,34), outcome entropy(35), and instrumental divergence
406 (36,49). Further experiments are needed to understand how these variables interact with personal
407 control in the acquisition of potential control over the environment.

408 Interestingly, when subjects selected the *free* option, the subsequent choice was sensitive
409 to the past reward when the terminal state (the selected target) was coherent and the reward could
410 therefore be attributed to the subject's action. In contrast, subjects' choices were insensitive to past
411 reward when the terminal state was incoherent. Furthermore, the probability of sticking with the
412 previous 2nd-stage choice following incoherent trials, whether rewarded or not, was not different
413 from the probability of sticking with the previously *rewarded* 2nd-stage choice following coherent
414 trials. Thus, subjects appeared to ignore information about action-state-reward contingencies that
415 was externally derived, and instead appeared to double down by repeating their past choice as if
416 they sought to maintain or regain personal control. This behavior is consistent with many
417 observations suggesting that when individuals experience situations that threaten or reduce their
418 personal control, they implement compensatory strategies to restore their perceived control to its
419 baseline level(50,51).

420 Computationally, however, this compensatory strategy is at odds with a pure model-based
421 strategy(37), where an agent could exploit information about action-state-reward contingencies
422 whether it derived from their own choices (internal control) or from the environment (external
423 control). Rather, it is consistent with work showing that choice-seeking could emerge when self-
424 determined actions amplify subsequent positive reward prediction errors(5,52), and more generally
425 with the notion that events are processed differently depending on individuals' beliefs about their
426 own control abilities. Thus, positive events are amplified only when they are believed to be within
427 one's personal control, whereas they are treated impartially when they are not(52), or when they
428 come from an uncontrollable environment(53).

429 Together, our results suggest that choice seeking may represent one critical facet of intrinsic
430 motivation and is associated with the desire of personal control. They also suggest that the need for
431 personal control can compete with maximization of extrinsic reward provided by externally driven
432 actions. Indeed, subjects favor positive outcomes associated to internally driven action even if
433 reward rate is lower than for action performed under the instruction of an external agent. In general,
434 the perception of being in personal control could then account for several aspects of our daily life
435 such as enjoyment during game(54) or motivation to perform demanding task(55). Since our results
436 shown inter-individual difference, it would be nonetheless important in the future to phenotype
437 subject behaviors during choice-making to investigate how these individual traits can explain
438 attitude difference when facing decision and their consequences, as exemplified by the variety of
439 attribution and explanation styles of individuals in the general population(56,57).

440

441

442

443

444 **Materials and Methods:**

445 **Participants.** Ninety-four healthy individuals (mean age = 30 \pm SD 7.32 years, 64 females)
446 responded to posted advertisements and were recruited to participate in this study. Relevant
447 inclusion criteria for all participants were being fluent in French, not treated for neuropsychiatric
448 disorders, having no color vision deficiency and being aged between 18 and 45 years old. Out of
449 these 94 subjects, 58 participated to experiment 1 and 36 to experiments 2-3. We gave subjects 40
450 euros for participating. The sample size was chosen based on previous studies that used similar
451 two-alternative decision making tasks(52,58,59).

452

453 **Ethics statement.** The local ethics committee (Comité d’Evaluation Éthique de l’Inserm) approved
454 the study (2019-CER2-MR-004). Participants gave written informed consent during inclusion in
455 the study, which was carried out in accordance with the declaration of Helsinki (1964; revised
456 2013).

457

458 **General procedure.** The paradigm was written in Matlab, using the Psychophysics Toolbox
459 extensions(60,61). It was presented on a 24 inches screen (1920 x 1080 pixels, aspect ratio 16:9).
460 Subjects seat \sim 57 cm from the center of the monitor. Our behavioral task design was designed as a
461 value-based decision paradigm. All participants received written and oral instructions. They were
462 told that the goal of the task was to gain the maximum number of rewards (a large green euro).
463 They were informed about the differences between the different trial types and that the extrinsic
464 reward contingencies experienced during the training phases remained identical during the test
465 phases. After instructions, participants received a pre-training session of a dozen trials (pre-train
466 and pre-test phases) in order to familiarize them with the task design and the keys they would have

467 to press.

468 In our experiments, subjects performed repeated trials with a two-stage structure. In the 1st-
469 stage they made an initial decision about what could occur in the 2nd-stage. Selecting the *free*
470 option led to a subsequent opportunity to choose and selecting the *forced* option led to an obligatory
471 computer-selected action. In the 2nd-stage, we presented subjects with two fractal images, one of
472 them being more rewarded following *free* selection in experiment 1 (except for P=0.5) and
473 experiment 2. In experiments 1 and 2, the computer always selected the same fractal target
474 following *forced* selection. Experiment 3 all fractal targets were equally rewarded and the computer
475 randomly selected one of the two fractal targets following *forced* selection (50%). Following *forced*
476 selection, the target to select with a key press was indicated by a grey V-shape above the target.
477 Pressing the other key on this trial type did nothing and the computer waited for the correct key
478 press to proceed further in the trial sequence. Either at the 1st- or 2nd-stage, after the subject's
479 selection of the target, a red V-shape appears immediately after above the target to indicate the one
480 they had selected (in experiment 3 blocks this red V-shape remains 250ms on the screen and
481 eventually jumps with the target, see below).

482

483 **Experimental conditions.** In experiment 1, fifty-eight subjects performed trials where the
484 maximal reward probabilities were matched following *free* and *forced* selection. We varied the
485 overall expected value across different blocks of trials, each of them being associated to different
486 programmed extrinsic reward probabilities (P). Forty-eight subjects performed a version with 3
487 blocks (experiment 1a) with different extrinsic reward probabilities ranging from 0.5 to 1 (block 1:
488 $P_{forced} = P_{free} = 0.5$; block 2: $P_{forced} = 0.75$, $P_{free}|a1 = 0.75$, $P_{free}|a2 = 0.25$; block 3: $P_{forced} = 1$, $P_{free}|$
489 $a1 = 1$, $P_{free}|a2 = 0$; where a1 and a2 represent the two possible key presses associated with the
490 fractal targets). Ten additional subjects performed the same task with 4 different blocks

491 (experiment 1b) associated to extrinsic reward probabilities also ranging from 0.5 to 1 ($P = 0.5$ or
492 0.67 or 0.83 or 1 from block 1 to 4 respectively.) We did not observe any substantial difference
493 between these two subject groups, and pooled them for analyses.

494 Experiment 2 was similar to experiment 1 (six states) except programmed extrinsic reward
495 associated with the *forced* option were higher than than the *free* option in two out of three blocks
496 ($P_{forced} = 0.75, 0.85$ or 0.95). Reward probabilities following *free* selection did not change across
497 the three blocks ($P_{free}|a1 = 0.75, P_{free}|a2 = 0.25$)

498 Experiment 3 consisted of a 7-state version of the two-stage task. Here, we manipulated the
499 coherence between the subject selection of a 2nd-stage (fractal) target and the target ultimately
500 displayed on the screen by the computer. Irrespectively of the target finally selected by the
501 computer or the subjects, the extrinsic reward probability associated to all the 2nd-stage targets in
502 *free* and *forced* trials was set at $P=0.75$. Importantly, adding the 7th state in this last task version
503 allowed the computer to swap the fractal 2nd-stage targets following both *free* and *forced* selection.
504 Thus, subjects did not perceive the weak coherence as a feature specific to the *free* condition.

505 We associated unique fractal targets with each action in the 2nd-stage, and a new set was
506 used for each block in all experiments. Colors of the 1st-stage targets were different between
507 experiments. Positive or negative reward feedback, as well as the side of the 1st-stage and 2nd-stage
508 target positions, were pseudo-randomly interleaved on the right or left of screen center. Feedback
509 was represented by the presentation (reward) or not (non-reward) of a large green euro image.

510 In experiment 1, when $P<1$, participants performed a minimum of 48 trials per block in the
511 training phases (*forced* and *free*) and the test phases. For $P=1$, participants performed a minimum
512 24 trials for training phases (*forced* and *free*) and 48 trials for test phase. The order of the blocks
513 were randomly interleaved. In experiments 2 and 3 they performed a minimum of 40 trials for each
514 block. Here, subjects started by performing experiment 3 followed by experiment 2. This was to

515 ensure that the value of *free* trials was not devalued by experiment 2 (titration) when performing
516 experiment 3. In experiment 3, subjects always started by the block with no target swaps
517 (incoherence = 0), and in experiment 2 by the block with equal extrinsic reward probability
518 (equivalent to the block P=0.75 of experiment 1). All the other blocks were randomly interleaved.

519

520 **Trial structure.** During the training phase, for each trial, a first fixation point appeared in the
521 center of the screen for 500ms, followed by the one of the first two targets of the different trial
522 types for an additional 750ms, either (*forced* or *free*) to the left or right of the fixation point (~11°
523 from the center of the screen on the horizontal axis, 3° wide). Immediately after, the first target
524 was turned off and two fractal targets appeared at the same eccentricity than the first target to the
525 left and right of the fixation point. The subjects could then choose by themselves or had to match
526 the target (depending on the trial type) using a key press (left or right arrow keys for left and right
527 targets, respectively). After their selection, a red V-shape appeared for about 1000ms above the
528 selected target (trace epoch). Note that in experiment 3, the V-shape was initially light red and
529 turned on for 250ms above the actual fractal target selected by the subjects. It was then turn in dark
530 red for 750ms. If the trial was incoherent, after 250ms, the red V-shape jumped and thus reappeared
531 simultaneously with the other target on the other side of the screen also for 750ms. Finally, the
532 fixation point was turned-off and the outcome was displayed during 750ms before the next trial.
533 For the test phase, the timing was equivalent except for the decision epoch related to the first stage
534 where participants could choose their favorite trial type (*free* and *forced* targets positioned
535 randomly, left or right) after 500ms of fixation point presentation. When a selection was made, the
536 first target remained for 500ms, associated to a red V-shape over the selected 1st-stage target –
537 indicating their choice. The second stage started with a 500ms epoch where only the fixation point
538 was presented on the screen, followed by the fractal target presentation. During the first and second

539 action epochs, no time pressure was imposed on subjects to make their choice, but if they pressed
540 one of the keys during the first 100ms after target presentation ('early press'), a large red cross was
541 displayed in the center of the screen for 500ms and the trial was repeated.

542

543 **Computational modelling.** We fitted individual subject data with variants of temporal-difference
544 reinforcement learning (TDRL) models. All models maintained a look-up table of state-action
545 value estimates ($Q(s, a)$) for each unique target and each action across all conditions within a
546 particular experiment. State-action values were updated at each stage ($t \in \{1, 2\}$) within a trial
547 according to the prediction error measuring the discrepancy between obtained and expected
548 outcomes:

$$549 \quad \delta_t = r_{t+1} + Z(s_{t+1}, a_{t+1}) - Q(s_t, a_t)$$

550 where $r_{t+1} \in \{0, 1\}$ indicates whether the subject received an extrinsic reward, and $Z(s_{t+1}, a_{t+1})$
551 represents the current estimate of the state-action value. The latter could take three possible forms:

$$552 \quad Z(s_{t+1}, a_{t+1}) = \begin{cases} Q(s_{t+1}, a_{t+1}) & \text{SARSA} \\ \max_{a'} Q(s_{t+1}, a') & \text{Q-learning} \\ \beta \cdot \max_{a'} Q(s_{t+1}, a') + (1 - \beta) \cdot \min_{a'} Q(s_{t+1}, a') & \beta\text{-pessimistic} \end{cases}$$

553 Although Q-learning and SARSA variants differ in whether they learn off- or on-policy,
554 respectively, we treated both of these algorithms as optimistic. Q-learning is strictly optimistic by
555 considering only the best future state-action value, whereas SARSA can be more or less optimistic
556 depending on the sensitivity of the mapping from state-action value differences to behavioral
557 policy. We compared Q-learning and SARSA variants with a third state-action value estimator that
558 incorporates risk attitude through a weighted mixture of the best and worst future action values
559 (Gaskett's β -pessimistic model(32)). As $\beta \rightarrow 1$ the pessimistic estimate of the current state-action
560 value converges to Q-learning.

561 The prediction error was then used to update all state-action values according to:

562
$$Q(s_{t+1}, a_{t+1}) \leftarrow Q(s_{t+1}, a_{t+1}) + \alpha \cdot \delta_t$$

563 where $\alpha \in [0,1]$ represents the learning rate.

564 We tested whether a free choice bonus could explain choice preference by modifying the
565 obtained reward as follows:

566
$$r_{t+1} = r_{t+1}^{\text{extrinsic}} + \rho$$

567 where $\rho \in (-\infty, +\infty)$ is a scalar parameter added to any extrinsic reward following any action
568 taken following selection of the free option.

569 Free actions at each stage were generated using a softmax policy as follows:

570
$$\pi(s, a^1) = \frac{\exp(Q(s, a^1)/\tau)}{\exp(Q(s, a^1)/\tau) + \exp(Q(s, a^2)/\tau)}$$

571 where increasing the temperature, $\tau \in [0, +\infty)$, produces a softer probability distribution over
572 actions. The forced option, on the other hand, always led to the same fixed action. We used a
573 softmax behavioral policy for all TDRL variants, and in the context of our task, the Q-learning and
574 SARSA algorithms were often similar in explaining subject data, so we treated them together in
575 the main text (Supplementary Fig. 1).

576 We also tested the possibility that subjects exhibited tendencies to alternate or persevere
577 following free or forced actions. We implemented this using a stickiness parameter that modified
578 the policy as follows:

579
$$\pi(s, a^1) = \frac{\exp[(Q(s, a^1) + \kappa \cdot C_t(s, a^1))/\tau]}{\exp[(Q(s, a^1) + \kappa \cdot C_t(s, a^1))/\tau] + \exp[(Q(s, a^2) + \kappa \cdot C_t(s, a^2))/\tau]}$$

580 where the $\kappa \in (-\infty, +\infty)$ parameter represents the subject's tendency to persevere, and $C_t(s, a)$
581 is a binary indicator for which fractal and action was chosen on the previous trial.

582 We independently combined the free parameters to produce a family of model fits for each

583 subject. We allowed the learning rate (α) and softmax temperature (τ) to differ for each of the two
584 stages in a trial. We therefore fitted a total of 48 models (3 estimates of current state-action value
585 [SARSA, Q, β -pessimistic] \times presence or absence of free choice bonus [ρ] \times 2- vs 1-learning rate
586 [α] \times 2- vs 1-temperature [τ] \times presence or absence of stickiness [κ]).

587

588 **Parameter estimation and model comparison.** We fitted model parameters using maximum a
589 posteriori (MAP) estimation using the following priors:

590 $\alpha \sim \text{beta}(\text{shape1}=1.1, \text{shape2}=1.1)$

591 $1/\tau \sim \text{gamma}(\text{shape}=1.2, \text{scale}=5)$

592 $\beta \sim \text{beta}(\text{shape1}=1.1, \text{shape2}=1.1)$

593 $\rho \sim \text{norm}(\text{mean}=0, \text{sd}=1)$

594 $\kappa \sim \text{norm}(\text{mean}=0, \text{sd}=1).$

595 We based hyperparameters for α and $1/\tau$ on Daw and colleagues (37). We used the same priors
596 and hyperparameters for all models containing a particular parameter. We used limited-memory
597 quasi-Newton algorithm (L-BFGS-B) to numerically compute MAP estimates, with α and β
598 bounded between 0 and 1 and $1/\tau$ bounded below at 0. For each model, we selected the best MAP
599 estimate from 10 random parameter initializations.

600 For each model for each subject, we fitted a single set of parameters to both training and
601 test data across conditions. We initialized state-action values to zero at the beginning of the training
602 phase for each condition. Data from the training phase consisted of 2nd-stage actions and rewards,
603 but we also presented subjects with the 1st-stage cues corresponding to the condition being trained
604 (forced or free). Therefore, we fitted the TDRL models assuming that the state-action values
605 associated with the 1st-stage fractals also underwent learning during the training phase, and that

606 these backups continued into the test phase, where subjects actually made 1st-stage decisions. That
607 is, we initialized the state-action values during the test phase with the final state-action values
608 during the training phase.

609 We used Schwarz weights to compare models, which provides a measure of the strength of
610 evidence in favor of one model over others and can be interpreted as the probability that a model
611 is best in the Bayesian Information Criterion (BIC) sense(62). We calculated weights for each
612 model as:

613
$$w_i(\text{BIC}) = \frac{\exp(-\Delta_i(\text{BIC})/2)}{\sum_{k=1}^K \exp(-\Delta_k(\text{BIC})/2)}$$

614 so that $\sum w_i(\text{BIC}) = 1$. We selected the model with the maximal Schwarz weight for each subject.

615 In order to verify that we could discriminate different state-action value estimates and how
616 accurately we could estimate parameters, we performed model and parameter recovery analyses on
617 simulated datasets (Supplementary Fig. 1).

618

619 **Statistical analyses.** We used generalized linear mixed models (GLMM) to examine differences
620 in choice behavior. When the model did not include trial-specific information (e.g., reward on the
621 previous trial), we aggregated data to the block level. Otherwise, we used choice data at the trial
622 level. We included random effects by subject for all models (random intercepts and random slopes
623 for the variable manipulated in each experiment; maximal expected value, relative expected value,
624 or incoherence for experiments 1, 2, and 3, respectively). We performed GLMM significance
625 testing using likelihood-ratio tests, and we corrected for multiple comparisons in post-hoc tests
626 using Tukey's method. We used generalized additive mixed models (GAMM) to examine choice
627 behavior as a function of trial within a block. We obtained smooth estimates of choice behavior
628 using penalized regression splines, with penalization that allowed smooths to be reduced to zero

629 effect(63). We included separate smooths by block. We performed GAMM significance testing
630 using approximate Wald-like tests(64).

631

632 **References**

633

- 634 1. Leotti LA, Iyengar SS, Ochsner KN. Born to choose: The origins and value of the need for
635 control. *Trends in Cognitive Sciences*. 2010.
- 636 2. Suzuki S. Effects of number of alternatives on choice in humans. *Behav Processes*. 1997;
- 637 3. Bown NJ, Read D, Summers B. The Lure of Choice. *J Behav Decis Mak*. 2003;
- 638 4. Leotti LA, Delgado MR. The inherent reward of choice. *Psychol Sci*. 2011;
- 639 5. Cockburn J, Collins AGE, Frank MJ. A Reinforcement Learning Mechanism Responsible for
640 the Valuation of Free Choice. *Neuron*. 2014;
- 641 6. Bobadilla-Suarez S, Sunstein CR, Sharot T. The intrinsic value of choice: The propensity to
642 under-delegate in the face of potential gains and losses. *J Risk Uncertain*. 2017;
- 643 7. Voss SC, Homzie MJ. Choice as a Value. *Psychol Rep*. 1970;
- 644 8. Catania AC. FREEDOM AND KNOWLEDGE: AN EXPERIMENTAL ANALYSIS OF PREFERENCE
645 IN PIGEONS 1 . *J Exp Anal Behav*. 1975;
- 646 9. Suzuki S. Selection of forced- and free-choice by monkeys (*Macaca fascicularis*). *Percept
647 Mot Skills*. 1999;
- 648 10. Perdue BM, Evans TA, Washburn DA, Rumbaugh DM, Beran MJ. Do monkeys choose to
649 choose? *Learn Behav*. 2014;
- 650 11. Beattie J, Baron J, Hershey JC, Spranca MD. Psychological determinants of decision
651 attitude. *J Behav Decis Mak* [Internet]. 1994 Jun 1 [cited 2022 Jun 13];7(2):129–44.
652 Available from: <https://onlinelibrary.wiley.com/doi/full/10.1002/bdm.3960070206>
- 653 12. Ly V, Wang KS, Bhanji J, Delgado MR. A reward-based framework of perceived control.
654 *Front Neurosci*. 2019;13(FEB):65.
- 655 13. Ryan RM, Deci EL. Self-determination theory and the facilitation of intrinsic motivation,
656 social development, and well-being. *Am Psychol*. 2000;55(1):68–78.
- 657 14. Deci EL, Ryan RM. Intrinsic Motivation and Self-Determination in Human Behavior.
658 *Intrinsic Motivation and Self-Determination in Human Behavior*. 1985.
- 659 15. Bandura A, Freeman WH, Lightsey R. Self-Efficacy: The Exercise of Control. *J Cogn
660 Psychother* [Internet]. 1999 Jan 1 [cited 2022 Jun 13];13(2):158–66. Available from:
661 <https://connect.springerpub.com/content>
- 662 16. Maier SF, Seligman MEP. Learned Helplessness: Theory and Evidence. *J of Exp Psychol
663 Gen*. 1976;105(1):3–46.
- 664 17. deCharms R. Personal causation: The internal affective determinants of behavior. New
665 York: Academic Press; 1968. 1–398 p.
- 666 18. White RW. Motivation reconsidered: The concept of competence. *Psychol Rev* [Internet].
667 1959 Sep [cited 2022 Jun 13];66(5):297–333. Available from: /record/1961-04411-001
- 668 19. Ajzen I. Perceived behavioral control, self-efficacy, locus of control, and the theory of

669 planned behavior. *J Appl Soc Psychol.* 2002;

670 20. Hull CL. *Principles of behavior*. New York: Appleton-Century-Crofts; 1943.

671 21. Bromberg-Martin ES, Monosov IE. Neural circuitry of information seeking. *Curr Opin*
672 *Behav Sci.* 2020 Oct 1;35:62–70.

673 22. Kidd C, Hayden BY. The Psychology and Neuroscience of Curiosity. *Neuron.* 2015 Nov
674 4;88(3):449–60.

675 23. Thrun SB. *Efficient Exploration In Reinforcement Learning*. Pittsburgh: Carnegie Mellon
676 University; 1992.

677 24. Fowler H. *Curiosity and Exploratory Behavior*. New York: Macmillan; 1965.

678 25. Brehm JW. Postdecision changes in the desirability of alternatives. *J Abnorm Soc Psychol.*
679 1956 May;52(3):384–9.

680 26. Festinger L. *A Theory of Cognitive Dissonance [Internet]*. Stanford: Stanford UP; 1957
681 [cited 2022 Jun 13]. Available from: https://books.google.fr/books?hl=fr&lr=&id=voeQ-8CASacC&oi=fnd&pg=PA1&ots=9z87Msw9uB&sig=YErlqdxMzgp8ZeMa0i55CPXm3w&redir_esc=y#v=onepage&q&f=false

682 27. Sharot T, Velasquez CM, Dolan RJ. Do decisions shape preference? Evidence from blind
683 choice. *Psychol Sci [Internet]*. 2010 Sep [cited 2022 Jun 13];21(9):1231. Available from:
684 [/pmc/articles/PMC3196841/](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3196841/)

685 28. Izuma K, Matsumoto M, Murayama K, Samejima K, Sadato N, Matsumoto K. Neural
686 correlates of cognitive dissonance and choice-induced preference change. *Proc Natl Acad
687 Sci.* 2010;107:22014–9.

688 29. Smith JE, Winkler RL. The Optimizer’s Curse: Skepticism and Postdecision Surprise in
689 Decision Analysis. *Manage Sci [Internet]*. 2006 Mar 1 [cited 2022 Jun 13];52(3):311–22.
690 Available from: <https://pubsonline.informs.org/doi/abs/10.1287/mnsc.1050.0451>

691 30. Hasselt H. Double Q-learning. Vol. 23, *Advances in neural information processing systems*.
692 New York: Macmillan; 2010.

693 31. Zorowitz S, Momennejad I, Daw ND. Anxiety, Avoidance, and Sequential Evaluation.
694 *Comput Psychiatry.* 2020;

695 32. Gaskett C. Reinforcement learning under circumstances beyond its control. In:
696 Proceedings of the International Conference on Computational Intelligence for Modelling
697 Control and Automation [Internet]. 2003 [cited 2022 Jun 13]. Available from:
698 <http://www.his.atr.co.jp/~cgaskett/>

699 33. Ayal S, Zakay D. The perceived diversity heuristic: the case of pseudodiversity. *J Pers Soc
700 Psychol [Internet]*. 2009 Mar [cited 2022 Jul 27];96(3):559–73. Available from:
701 <https://pubmed.ncbi.nlm.nih.gov/19254103/>

702 34. Schwartenbeck P, Fitzgerald THB, Mathys C, Dolan R, Kronbichler M, Friston K. Evidence
703 for surprise minimization over value maximization in choice behavior. *Sci Rep [Internet]*.
704 2015 Nov 13 [cited 2022 Jul 27];5. Available from:
705 <https://pubmed.ncbi.nlm.nih.gov/26564686/>

706 35. Erev I, Barron G. On adaptation, maximization, and reinforcement learning among
707 cognitive strategies. *Psychol Rev [Internet]*. 2005 Oct [cited 2022 Jul 27];112(4):912–31.
708 Available from: <https://pubmed.ncbi.nlm.nih.gov/16262473/>

709 36. Mistry P, Liljeholm M. Instrumental Divergence and the Value of Control. *Sci Reports*
710 2016 61 [Internet]. 2016 Nov 8 [cited 2022 Jul 26];6(1):1–10. Available from:
711 <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5107000/>

712

713 https://www.nature.com/articles/srep36295
714 37. Daw ND, Gershman SJ, Seymour B, Dayan P, Dolan RJ. Model-based influences on
715 humans' choices and striatal prediction errors. *Neuron* [Internet]. 2011/03/26.
716 2011;69(6):1204–15. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/21435563>
717 38. Sharot T, Martino B De, Dolan RJ. How Choice Reveals and Shapes Expected Hedonic
718 Outcome. *J Neurosci* [Internet]. 2009 Mar 25 [cited 2022 Jun 13];29(12):3760–5. Available
719 from: <https://www.jneurosci.org/content/29/12/3760>
720 39. Wang KS, Kashyap M, Delgado MR. The Influence of Contextual Factors on the Subjective
721 Value of Control. *Emotion* [Internet]. 2021 [cited 2022 Jul 4];21(4):881–91. Available
722 from: <https://dx.doi.org/10.1037/emo0000760>
723 40. Chew SH, Ho JL. Hope: An empirical study of attitude toward the timing of uncertainty
724 resolution. *J Risk Uncertain*. 1994;
725 41. Ahlbrecht M, Weber M. The Resolution of Uncertainty: An Experimental Study. *J
726 Institutional Theor Econ JITE*. 1996;
727 42. Zheng Z, Oh J, Hessel M, Xu Z, Kroiss M, Van Hasselt H, et al. What can learned intrinsic
728 rewards capture? In: 37th International Conference on Machine Learning, ICML 2020.
729 2020.
730 43. Singh S, Lewis RL, Barto AG, Sorg J. Intrinsically Motivated Reinforcement Learning: An
731 Evolutionary Perspective. *IEEE Trans Auton Ment Dev*. 2010;
732 44. Botvinick M, Ritter S, Wang JX, Kurth-Nelson Z, Blundell C, Hassabis D. Reinforcement
733 Learning, Fast and Slow. *Trends in Cognitive Sciences*. 2019.
734 45. Maier SF, Seligman MEP. Learned helplessness at fifty: Insights from neuroscience.
735 *Psychol Rev*. 2016 Jul 1;123(4):349–67.
736 46. Haggard P, Chambon V. Sense of agency. *Curr Biol* [Internet]. 2012 May 22 [cited 2022
737 Jun 13];22(10). Available from: <https://pubmed.ncbi.nlm.nih.gov/22625851/>
738 47. Farkas BC, Chambon V, Jacquet PO. Do perceived control and time orientation mediate
739 the effect of early life adversity on reproductive behaviour and health status? Insights
740 from the European Value Study and the European Social Survey. *Humanit Soc Sci
741 Commun* 2022 91 [Internet]. 2022 Feb 14 [cited 2022 Jun 13];9(1):1–14. Available from:
742 <https://www.nature.com/articles/s41599-022-01066-y>
743 48. Kraus MW, Piff PK, Mendoza-Denton R, Rheinschmidt ML, Keltner D. Social class,
744 solipsism, and contextualism: How the rich are different from the poor. *Psychol Rev*.
745 2012;119(3):546–72.
746 49. Liljeholm M. Instrumental Divergence and Goal-Directed Choice. In: *Goal-Directed
747 Decision Making* [Internet]. Academic Press; 2018 [cited 2022 Jul 27]. p. 27–48. Available
748 from: <https://doi.org/10.1016/B978-0-12-812098-9.00002-4>
749 50. Landau MJ, Kay AC, Whitson JA. Compensatory control and the appeal of a structured
750 world. *Psychol Bull* [Internet]. 2015 May 1 [cited 2022 Jun 13];141(3):694–722. Available
751 from: <https://pubmed.ncbi.nlm.nih.gov/25688696/>
752 51. Whitson JA, Galinsky AD. Lacking control increases illusory pattern perception. *Science
753 (80-)* [Internet]. 2008 Oct 3 [cited 2022 Jun 30];322(5898):115–7. Available from:
754 <https://www.science.org/doi/10.1126/science.1159845>
755 52. Chambon V, Théro H, Vidal M, Vandendriessche H, Haggard P, Palminteri S. Information
756 about action outcomes differentially affects learning from self-determined versus

757 imposed choices. *Nat Hum Behav*. 2020;

758 53. Dorfman HM, Bhui R, Hughes BL, Gershman SJ. Causal Inference About Good and Bad
759 Outcomes. *Psychol Sci* [Internet]. 2019 Apr 1 [cited 2022 Jun 13];30(4):516–25. Available
760 from: <https://journals.sagepub.com/doi/full/10.1177/0956797619828724>

761 54. Hulaj R, Nyström MBT, Sörman DE, Backlund C, Röhlcke S, Jonsson B. A Motivational
762 Model Explaining Performance in Video Games. *Front Psychol* [Internet]. 2020 Jul 14
763 [cited 2022 Jun 13];11. Available from: <https://pubmed.ncbi.nlm.nih.gov/32760321/>

764 55. Sidarus N, Palminteri S, Chambon V. Cost-benefit trade-offs in decision-making and
765 learning. *PLoS Comput Biol* [Internet]. 2019 [cited 2022 Jun 13];15(9). Available from:
766 <https://pubmed.ncbi.nlm.nih.gov/31490934/>

767 56. Rotter JB. Generalized expectancies for internal versus external control of reinforcement.
768 *Psychol Monogr* [Internet]. 1966 [cited 2022 Jun 14];80(1):1–28. Available from:
769 [/record/2011-19211-001](https://doi.org/10.1037/0033-295X.80.1.1)

770 57. Abramson LY, Seligman ME, Teasdale JD. Learned helplessness in humans: Critique and
771 reformulation. *J Abnorm Psychol* [Internet]. 1978 Feb [cited 2022 Jun 14];87(1):49–74.
772 Available from: [/record/1979-00305-001](https://doi.org/10.1037/0021-9010.87.1.49)

773 58. Palminteri S, Lefebvre G, Kilford EJ, Blakemore SJ. Confirmation bias in human
774 reinforcement learning: Evidence from counterfactual feedback processing. *PLOS Comput
775 Biol* [Internet]. 2017 Aug 1 [cited 2022 Jun 16];13(8):e1005684. Available from:
776 <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005684>

777 59. Palminteri S, Khamassi M, Joffily M, Coricelli G. Contextual modulation of value signals in
778 reward and punishment learning. *Nat Commun* [Internet]. 2015 Aug 25 [cited 2022 Jun
779 16];6. Available from: <https://pubmed.ncbi.nlm.nih.gov/26302782/>

780 60. Brainard DH. The Psychophysics Toolbox. *Spat Vis*. 1997;

781 61. Kleiner M, Brainard DH, Pelli DG, Broussard C, Wolf T, Niehorster D. What's new in
782 Psychtoolbox-3? *Perception*. 2007;

783 62. Wagenmakers EJ, Farrell S. AIC model selection using Akaike weights. *Psychon Bull Rev*
784 2004 111 [Internet]. 2004 [cited 2022 Jun 14];11(1):192–6. Available from:
785 <https://link.springer.com/article/10.3758/BF03206482>

786 63. Wood SN. Generalized Additive Models. An Introduction with R, Second Edition.
787 Chapman and Hall; 2017. 496 p.

788 64. Wood SN. On p-values for smooth components of an extended generalized additive
789 model. *Biometrika* [Internet]. 2013 Mar 1 [cited 2022 Jul 7];100(1):221–8. Available from:
790 <https://academic.oup.com/biomet/article/100/1/221/192816>

791

792

793

794 **Acknowledgements:** J.M. was supported by the Agence Nationale de la Recherche (ANR) grant
795 ANR-19-CE37-0014-01 (ANR PRC) and by the European Commission (H2020-MSCA-IF-2018-
796 #845176). D.B. was supported by a FRM fellowship (FDM201906008526). V.C. was supported
797 by the ANR grants ANR-17-EURE-0017 (Frontiers in Cognition), ANR-10-IDEX-0001-02 PSL

798 (program ‘Investissements d’Avenir’), ANR-16-CE37-0012-01 (ANR JCJ) and ANR-19-CE37-
799 0014-01 (ANR PRC). B.L. was supported by the ANR grant ANR-19-CE37-0014-01. The authors
800 of this article are grateful to Karim Ndiaye, operational manager of the PRISME platform at the
801 ICM for his valuable help during participant testing.

802
803 **Author Contributions:** J.M., V.C. and B.L. designed the study; J.M., M.R.A., D.B. and A.K.
804 performed the experiments and preliminary analyses V.C.; J.M., and B.L. designed and performed
805 final analyses; J.M., V.C. and B.L. wrote the manuscript.

806
807 **Data availability statement:** All data and related metadata underlying the findings reported will
808 be deposited in Zenodo (DOI: 10.5281/zenodo.7057043) at the time of publication.

809
810 **Code reporting:** Code written in support of this publication will be made publicly available in
811 Zenodo (DOI: 10.5281/zenodo.7057080) at the time of publication.

812
813
814