Title: Dynamic refinement of behavioral structure mediates dopamine-dependent credit assignment

Dopamine initially reinforces spatially similar and temporally proximal actions to actions that trigger dopamine release, and drives a gradual refinement of the entire behavioral repertoire to home-in on reward-producing actions.

Authors: Jonathan C.Y. Tang[1], Vitor Paixao[2,3], Filipe Carvalho[2,4], Artur Silva[2], Andreas Klaus[2], Joaquim Alves da Silva[2], Rui M. Costa[1,2,5]*

Affiliations:

[1]Department of Neuroscience, Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY 10027, USA

[2]Champalimaud Neuroscience Programme, Champalimaud Research, Champalimaud Foundation Lisbon, Portugal

[3]Kinetikos, Coimbra, Portugal

[4]Open Ephys Production Site, Lisbon, Portugal

[5]Allen Institute, Seattle, WA 98109, USA

*Correspondence: rc3031@columbia.edu

Abstract

Animals exhibit a diverse behavioral repertoire when exploring new environments and can learn which actions or action sequences produce positive outcomes. Dopamine release upon encountering reward is critical for reinforcing reward-producing actions[1–3]. However, it has been challenging to understand how credit is assigned to the exact action that produced dopamine release during continuous behavior. We investigated this problem with a novel self-stimulation paradigm in which specific spontaneous movements triggered optogenetic stimulation of dopaminergic neurons. We uncovered that dopamine self-stimulation rapidly and dynamically changes the structure of the entire behavioral repertoire. Initial stimulations reinforced not only the stimulation-producing target action, but also actions similar to the target and actions that occurred a few seconds before stimulation. Repeated pairings led to gradual refinement of the behavioral repertoire leading animals to home in on the target action. Reinforcement of action sequences revealed further temporal dependencies of behavioral refinement. Action pairs that tend to be spontaneously separated by long time intervals promoted a stepwise credit assignment, with early refinement of actions most proximal to stimulation and subsequent refinement of more distal actions. Thus, a retrospective reinforcement mechanism promotes gradual refinement of the entire behavioral repertoire to assign credit to specific actions and action sequences that lead to dopamine release.

Main Text

Background

Animals spontaneously transition amongst a repertoire of movements when exploring new environments. Movements or movement sequences that produce positive outcomes are reinforced and increase in frequency to maximize the obtainment of those outcomes[4,5]. However, it is still not completely clear how animals assign credit to the exact action that produce reward in the context of a continuous behavioral space. This credit assignment problem[2,6–9] during spontaneous behavior poses at least two main challenges. First, it is unclear how animals come to preferentially perform a specific reward-producing action or action sequence above other possibilities in the behavioral repertoire. Second, it is unclear how animals derive contingency between a reward-producing action and reward if there can be variable delays between action performance and reward delivery.

Dopamine (DA) has been proposed to mediate credit assignment[6,10]. At the cellular level, DA can facilitate synaptic plasticity in corticostriatal synapses[11] within a critical time window that is behaviorally relevant[12–14]. Still, it is unknown how DA changes the dynamics of spontaneous behavior to mediate credit assignment. We therefore developed a paradigm to investigate how DA shapes the evolution of continuous behavior during action learning to gain insights into the process of credit assignment.

Conventional operant conditioning paradigms[5,15–19] have helped derive principles of behavioral reinforcement, but they are not ideal for studying action credit assignment. In general, such paradigms do not permit the clean isolation of actions as the trigger for reward versus particular

70 locations or objects. In such paradigms, animals are also required to perform a series of
71 consummatory actions, such as approaching and interacting with reward-delivering devices to
72 retrieve reward. These requirements make it difficult to investigate how credit is assigned to a
73 specific action or action sequence in the behavioral repertoire during continuous behavior.
74
75 Until recently, technological and conceptual limits have made it difficult to study how the entire
76 structure of continuous behavior evolves as naive animals come to associate specific action or
77 action sequences with reward. To address previous limitations, we developed a new approach to
78 study action credit assignment. This approach directly reinforces specific spontaneous action(s)
79 by triggering dopaminergic neuron (DA neuron) excitation and DA release upon action
80 performance. It combines wireless inertial sensors, unsupervised clustering of continuous
81 behavior[20,21] and optogenetics[22] into a closed-loop system linking specific action performance to
82 immediate phasic DA release (Methods; Fig. 1a-f). This paradigm permits action detection and
83 reinforcement without requiring an animal to approach or interact with a place/object/cue, or to
84 perform consummatory behavior. These combined features overcome the aforementioned
85 caveats associated with conventional paradigms.
86
87 *Rapid reinforcement of actions via closed-loop dopamine stimulation*
88 To implement the action detection component of the closed loop system, we first classified the
89 entire behavioral repertoire of individual mice[23] mice in a grey-walled open field using inertial
90 sensors and unsupervised affinity propagation clustering[20,21] (Fig. 1d). Self-paced behavior was
91 monitored using a novel, wireless inertial sensor system (WEAR; Methods) that allows minimal
92 movement restraints, high resolution behavior monitoring and fast data transmission to open-

93 source hardware and software for online experimentation (Fig. 1b, Extended Data Fig. 1a).
94 Affinity propagation clustering is particularly well suited to cluster an unknown number of
95 clusters[20], is computationally efficient[24], and easily outputs similarity between clusters.
96 Clustering begins by processing accelerometer and gyroscope data to extract 4 features
97 discriminating postural changes, movement momentum, head and head-body rotations, and total
98 body accelerations. Feature values from 300 ms long segments of behavior were discretized into
99 histograms, upon which pairwise similarity comparisons could be made using a Earth-Mover's
100 Distance (EMD)[25] metric. The similarity matrix of all possible pairwise comparisons were fed
101 into an unsupervised affinity propagation clustering algorithm[20] (Methods), identifying naturally
102 occurring repertoire of 300 ms long behavioral clusters[21], or "actions" (Fig. 1c,Extended Data
103 Fig. 1b). The choice of 300 ms long movements was informed by previous studies[21,26]. Using
104 these parameters, we identified over 30 clusters of spontaneous behavior per individual (34.3 +/-
105 2.1 and 35.6 +/- 2.5 total actions per ChR2-YFP and YFP mice, respectively; mean +/- standard
106 deviation,15 ChR2-YFP and 10 YFP mice). We chose particular clusters of actions to be
107 reinforced (hereby named target action A).
108
109 To implement closed-loop reinforcement, we used Cre-dependent AAV viruses (EF1a-DIO-
110 expression cassette) to express channelrhodopsin ChR2-YFP[22] or the control protein YFP
111 bilaterally in DA neurons of the ventral tegmental area (VTA) [27,28] of DAT-Cre mice (Fig.
112 1a,Extended Data Fig. 2a-c). Using the wireless inertial sensor, we tracked behavior
113 continuously in a white open field and used the similarity metric to match ongoing 300 ms
114 behavioral segments to exemplars representing each mouse's repertoire of actions (Fig. 1d-e).
115 Upon a match to a defined target action (target action A), a 25 hz, 600 ms long train of

116 optogenetic stimulation was delivered to DA neurons of the VTA parabrachial pigmented area
117 (PBP) (30-60 ms delay, Fig. 1e). These target action As were different for different animals, and
118 were dispersed across a behavioral space (Fig. 1g).To evaluate whether stimulation parameters
119 triggered DA release similar in magnitude to that triggered by sucrose reward in food restricted
120 mice, we delivered random optogenetic stimulations to ChR2-YFP- or YFP-expressing VTA DA
121 neurons while monitoring DA release with the GRAB rDA1m sensor [29] in both ventral and
122 dorsal striatum (Fig. 1f). We also measured DA release in the same animals upon delivery of
123 sucrose while they were food deprived. Sucrose presentation led to a sharp increase in DA
124 release in both ventral and dorsal striatum (Fig. 1f). Interestingly, optogenetic stimulation of DA
125 neurons in VTA with the parameters described above, resulted in a similar phasic increase in DA
126 not only in ventral striatum but also in dorsal striatum (Fig. 1f). This is consistent with emerging
127 evidence showing the existence of dorsal striatum-projecting VTA neurons[30,31]. Thus, our
128 optogenetic stimulation triggered DA release similar in decay and spatial localization to that
129 triggered by sucrose reward in food restricted mice (Fig. 1f), offering us a suitable approach to
130 interrogate how pairing DA release with specific action performance leads to credit assignment.
131
132 Closed loop reinforcement for a specific action occurred over a 3-day, 60-90 minute/session
133 protocol designed to probe both intra- and inter-session changes in behavior (Fig. 1h-m,
134 Extended Data Fig. 3). Optogenetic stimulation of VTA DA neurons upon execution of a
135 particular target action (action A) resulted in significant increase in the frequency of action A for
136 ChR2-YFP, but not YFP mice (Fig. 1h, Extended Data Fig.3b). Increased action A in ChR2-YFP
137 animals depends on optogenetic stimulation, as removal of closed-loop stimulations resulted in
138 progressive extinction of action A (Fig.3h, Extended Data Fig.3d). Resuming paired stimulation

139 led to rapid re-instatement of action A (Fig. 1h, Extended Data Fig.3c,e). Interestingly, during
140 extinction, ChR2-YFP animals kept performing exploratory unrewarded bursts of action A,
141 which could explain rapid reinstatement (Extended Data Fig. 3e,f). This paradigm revealed that
142 just a few pairings with DA leads to rapid reinforcement, as changes in multiple parameters
143 including decreased trigger latency, increased action A frequency and increased average
144 behavioral similarity towards action A become significant following 10-15 stimulations (Fig. 1i,
145 Extended Data Fig. 4a-b).
146
147 We next examined if only action A changed in frequency or if other non-stimulated actions also
148 changed with closed-loop reinforcement of action A. We calculated baseline-normalized
149 frequency of all actions in the repertoire and ordered them as a function of similarity to the target
150 action (Fig. 1j). Earth-Mover's Distance (EMD)[21,25] was used to measure each action exemplar's
151 similarity to the target exemplar (Methods), with lower EMD value indicating increased
152 similarity. Surprisingly, we observed that optogenetic stimulation resulted in a dramatic change
153 in the entire behavioral repertoire. We observed that early in training actions most similar to
154 target tended to also increase in frequency (Fig. 1j-l, Extended Data Fig. 4c) whereas actions
155 most dissimilar to target tended to decrease in frequency. Repeated pairing led to refinement of
156 the actions that were performed at high frequency, and by late stages action A became the
157 predominant action being performed, with a sharp drop-off of non-target action frequencies as
158 similarity to target decreased (Fig. 1k-l). Such effects were not observed in YFP controls
159 (Extended Data Fig. 4d-e). These data suggested that early reinforcement results in rapid
160 reshaping of the entire behavioral repertoire, biasing animals towards actions similar to the target

161 action, and continued pairing resulted in gradual refinement and assignment of credit to the
162 specific target action.
163
164 *Dynamics of behavioral refinement during reinforcement*
165 To better describe individual action dynamics during reinforcement, we categorized actions (511
166 actions, *n*=15 ChR2-YFP animals) by the trajectories of their changes in frequency throughout
167 learning (Methods). Three meaningful types of trajectories were categorized, comprising over
168 94% of all actions. These types were characterized by either initial increase that remained stable
169 (Sustained Increase), initial increases that decreased over time (Transient Increase) and initial
170 decreases that remained stable (Decreased) (Fig 1m, Extended Data Fig. 5-6). We again
171 confirmed that the frequency dynamics type of each particular action was related to its similarity
172 to target, regardless of whether actions were sorted based on their raw or percentile similarity
173 scores (Extended Data Fig. 6b-c). Actions most similar to target were predominately Sustained
174 Increase types, while moderately similar actions mostly comprised of Sustained Increase or
175 Transient Increase types and more dissimilar actions are more of the Decreased type (Extended
176 Data Fig. 6b-c). Taken together, these finer resolution analyses indicate again that the dynamics
177 of action frequency are related in great part to the similarity to target action.
178
179 *Reinforcement and refinement after reversal of action-reward contingencies*
180 Next, we asked if animals could follow changes in contingency between action and closed-loop
181 DA stimulation. We therefore chose a different action, action B, which is clearly distinct from
182 the action A for each animal (Methods, Fig. 2a, Extended Data Fig. 1c) and started delivering
183 DA neuron optogenetic stimulation after action B. Chosen action A/B pairs were relatively

184 dissimilar in the context of entire action similarity distributions (Fig. 2b). Upon reinforcement,
185 previously trained ChR2-YFP, but not YFP animals showed increased action B performance
186 over time (Fig. 2c-e, Extended Data Fig. 7). In contrast, action A frequency changes clearly
187 moved in the opposite direction from that of action B over time (Fig. 2c). Maintenance of action
188 B performance depended on continual reinforcement (Fig. 2c, Extended Data Fig. 7d-e). Similar
189 to action A, action B credit assignment unfolds by initially biasing the entire repertoire, i.e.,
190 increasing the frequency of similar actions and reducing the frequency of dissimilar actions. This
191 was again followed by gradually refining for action B relative to similar actions as pairing
192 progressed (Fig. 2d-e, Extended Data Fig. 7f). To confirm that action learning is contingent on
193 action B appearing before reinforcement, we subjected trained animals to a contingency
194 degradation protocol in which we delivered a similar number of random stimulations uncoupled
195 to action B performance. Action B performance decreased following contingency degradation
196 and could be re-instated upon resuming the action B-stimulation contingency (Fig. 2f, Extended
197 Data Fig. 7g). These experiments indicate that animals can follow changes in the contingency
198 between actions and DA release and assign credit to a new action through a similar process of
199 behavioral repertoire refinement.
200
201 Although animals show similar patterns of behavioral refinement for actions A and B, animals
202 that previously credited an action (action A) for DA release did initially respond to reinforcement
203 of a new action (action B) differently from naïve animals (Fig. 2g-j). Whereas naïve animals
204 responded to initial reinforcements for target action A by significantly increasing action A
205 performance relative to the non-target action B (Fig. 2g,i,left graph), animals with a history of
206 reinforcement on action A animals responded to initial reinforcements of action B by increasing

207 non-target action A performance (Fig. 2g,i,right graph). This trend reverses later such that target
208 action B becomes significantly increased over the non-target action A (Fig. 2g,i,right graph).
209 YFP control animals showed no such trends (Fig. 2h,j). Thus, DA reinforcement does not simply
210 reinforce the recently performed, temporally contiguous action, but trigger previously credited
211 actions in the face of a new action-reward contingency that is not yet learned. This suggest again
212 that animals learned the contingency between action performance and DA release.
213
214 *Temporal constraints of DA-dependent reinforcement*
215 The contingency degradation results above indicate that the temporal relation between target
216 action and DA phasic activity is important for reinforcement (Fig. 2e). Reinforcement is thought
217 to occur on behavior that precedes reward in time[10,12,14,19], and while temporal contiguity
218 between action and reinforcement has long been recognized[32–34], it is not clear how the position
219 of an action relative to the time of DA phasic activity influences its subsequent frequency. We
220 investigated if in addition to behavioral similarity, the temporal relationship between action and
221 stimulation influenced the dynamics of behavioral repertoire evolution during reinforcement and
222 credit assignment.
223
224 We observed that the median inter-target action interval decreased with stimulation in ChR2-
225 YFP mice (Fig. 3a,b). We therefore examined the distribution of the action dynamic types
226 categorized above (Sustained Increase, Transient Increase, Decreased) according to both an
227 action's similarity to target and the median time of that action's performance leading into target
228 during baseline, before reinforcement protocol began (Fig. 3c-e). Action dynamic types showed
229 distinct distribution patterns for these two dependent variables (similarity and time). Further,

230 these two dependent variables were not significantly collinear (Methods). Thus, action similarity
231 to target as well as baseline temporal proximity to target should together predict action dynamic
232 type upon reinforcement better than either factor alone. To test this idea, we performed
233 multinomial logistic regression to assess whether 1- or 2-factor models best fit the observed
234 dynamics pattern that an action would follow upon reinforcement (Fig. 3f,g). The two-factor
235 model outperformed either one-factor models, and prediction of action dynamics type with this
236 model was significantly above chance as assessed by precision-recall curves, which is suitable
237 for evaluating datasets with imbalanced categories[35] (Fig. 3g). The beta coefficients indicated
238 that increased similarity to target and decreased median time to target increases prediction of
239 Sustained Increase and Transient Increase dynamic types relative to Decreased types
240 (Supplementary Table). These results suggest that DA may reshape behavioral repertoire by
241 reinforcing not only actions similar to the target action but also actions that happen to be
242 performed temporally close to the reinforcer, as suggested before[10,12,14,19].
243
244 To more rigorously test whether DA reinforcement acts in a retrospective or prospective manner,
245 we increased the resolution of analysis by examining 1st order action transitions leading into and
246 out of stimulation (Fig. 3h-j). By focusing analysis on action transitions enriched within specific
247 1.2 second moving windows, one could distinguish more clearly behavior that occurred leading
248 up to, during, and after DA stimulation. Our analyses showed that action transitions enriched in
249 windows up to 1.2 seconds prior to stimulation onset, as well as during stimulation, are
250 reinforced early on (Fig. 3i). However, this did not occur to action transitions following
251 stimulation, suggesting an asymmetric process. Indeed, action transitions enriched in windows
252 leading into stimulation were also preferentially reinforced relative to those enriched in windows

253 after stimulation (Fig. 3j). Thus, DA stimulation promotes reinforcement of behaviors occurring
254 during stimulation and a few seconds before stimulation.
255
256 *Credit assignment for action sequences*
257 In the real world, when animals are spontaneously shifting between actions in their repertoire,
258 outcomes are often not the result of a single action but rather of a sequence of actions performed
259 at variable intervals. We therefore investigated the dynamics of reinforcement when the release
260 of DA is contingent upon the performance of a sequence of 2 actions (target action 1 and 2, T1
261 and T2). We applied closed loop optogenetics to ask whether naïve animals can learn a T1→T2
262 reinforcement rule, where the delays between T1 and T2 are governed by the spontaneous
263 behavior of the animals and not experimentally controlled (n=15 ChR2-YFP and 10 YFP mice,
264 Fig. 4a, Extended Data Fig. 2a,d-e, Extended Data Fig. 8-10). Various T1/T2 pairs were
265 sampled, with focus on sequences sharing general commonalities in movement order across
266 animals (Extended Data Fig. 1d,f-g). Overall, mice learned to increase the performance of a
267 sequence of two actions to obtain DA stimulation. Some animals showed a ChR2-dependent
268 increase in reinforcement within 5 sessions, but others experienced a lag in learning (Fig. 4b).
269 We hypothesized that this could relate to the initial time distance between T2 trigger and the
270 closest distal T1 (T1→T2 interval). Indeed, animals reinforced for action pairs with initially long
271 interval values tended to show slower learning curves (Fig. 4c-d). To capture a learning time
272 point whereby individuals reach similar rising phase in their respective learning curves, a
273 criterion frequency was set (Methods). 14 of 15 trained animals eventually reached criterion
274 (Fig. 4e; Extended Data Fig. 8a-c). Sequence performance depended on continuing DA
275 reinforcement (Fig. 4f,g). Learning was also revealed by decreases in the median T1→ T2 time

intervals (Fig.4h-i) and convergence of T1-to-T2 frequency ratio towards 1 (Fig. 4j). To quantify the specific credit assignment of T1 and T2 we used a refinement index that compares the median frequency of actions uniquely similar to T1 with those uniquely similar to T2, with the frequencies normalized by either that of T1 or T2 (Methods). Values lower than 1 indicate that the target actions are being performed even more frequently than similar actions, and thus indicate greater refinement (Methods). By the end of learning, T1 and T2 became credited as the reward-producing actions relative to their similar counterparts (Fig. 4k). YFP controls did not show any of these trends (Fig.4c-d,4g-h). Thus, closed loop reinforcement promoted learning of a two-action sequence rule in freely moving mice starting from a naïve state.

Importantly, the initial median T1→T2 interval performed by ChR2-YFP animals was inversely related to the eventual number of sessions required for each animal to reach criterion frequency (Fig. 4l). A sigmoidal curve was fit to the data, showing that animals with longer open field T1→T2 intervals beyond the sigmoidal midpoint tended to face sudden increase in sessions to reach criterion frequency (Fig. 4l). ChR2-YFP animals were divided according to the half-maximum point of the sigmoidal curve into 'Fast Learners' and 'Slow Learners'. Fast Learners quickly reached criterion frequency and low T1→T2 time intervals, whereas Slow Learners experienced a time lag in reaching criterion frequency and low T1→T2 intervals. Slow Learners tended to suddenly increase the frequency of sequence performance in sessions that showed a drop in the median T1→T2 interval to below 2-4 seconds (Fig.4d,h). In contrast, there was no stable sigmoidal relationship between T1-T2 action similarities and sessions to criterion frequency (Extended Data Fig. 8d). Thus, the initial median time distances between distal action

298 T1 and proximal action T2(which produced DA stimulation) modulated how fast animals learned
299 to effectively perform the reinforced action sequence.
300
301 If DA is acting retrospectively to reinforce actions performed earlier in time, we hypothesized
302 that the action most proximal to reinforcement, T2, should experience earlier refinement relative
303 to the distal action, T1. We again used the median target normalized frequencies of actions
304 uniquely related to T1 or T2 as refinement indices (Methods). Proximal T2 clearly refines
305 towards its most refined level earlier than the distal T1, at least in some animals (Fig. 5a). By
306 subtracting the area under the refinement curve for T1 from the curve for T2, one could calculate
307 differential refinement between the two actions. Positive values indicate refinement
308 preferentially favoring T2, and vice versa. A linear relationship was found between open field
309 median T1→T2 interval and differential refinement between T1 and T2 (Fig. 5b). This suggests
310 for longer T1→T2 median intervals, the proximal action T2 spends more sessions being more
311 refined than the distal action T1. In contrast, there was no significant linear relationship between
312 the initial intervals between the execution of the proximal action that led to reward and the next
313 initiation of the sequence (T2→T1) or of the similarity between T1 and T2 actions, and the
314 dynamics of differential refinement between T1 and T2 (Fig. 5b, right graph, Extended Data Fig.
315 9a).
316
317 We next investigated if the differential refinement between T1 and T2 was different for slow and
318 fast learners. We analyzed changes in T1-T2 refinement curves relative to 'Starting Points' at
319 which the refinement indices of T1 and T2 are most similar or are biased towards the distal T1
320 rather than the proximal T2 action (Methods). All Slow Learners showed a pattern where they

321 initially refine the repertoire of T2 from these Starting Points, and after reaching a maximum
322 Turning Point, they start showing a bias towards greater T1 refinement (Fig. 5c). Notably, by
323 these Turning Points the median intervals of T1→T2, but not T2→T1 events had decreased
324 significantly relative to initial values (Fig. 5d, Extended Data Fig. 9b). Therefore, the median
325 T1→T2 interval decrease occurred before a decrease in the interval to perform the next sequence
326 (T2→T1), which started decreasing after the Turning Point (Fig. 5e). Using these learning
327 landmarks, we asked more rigorously how animals homed in on T1 vs T2 over time (Fig. 5f,
328 Extended Data Fig.10a). We found that animals initially refined the action proximal to DA
329 stimulation (T2, between Starting Point and Turning Point), whereas T1 refinement occurred
330 several sessions later, after the Turning point (Fig. 5f, Extended Data Fig.10a). Indeed, the
331 Turning Point coincided with an increased probability of the T1 being found within 3.6 secs
332 before T2 and reinforcement (Fig. 5g-h). These results indicate that animals can assign credit to
333 sequences of actions that lead to reinforcement, following similar retrospective dynamics that
334 were observed for single actions, whereby the actions most proximal to reinforcement are refined
335 earlier and the actions more distal to reinforcement refined later, when they probabilistically start
336 to occur within a few seconds of DA release.
337
338 *Discussion*
339 Our results demonstrate that DA reinforcement promotes single action credit assignment from a
340 naïve state through a dynamic process whereby the entire behavioral repertoire is restructured.
341 During the initial stages of reinforcement both actions similar to the target action and actions that
342 were performed in close temporal proximity of the target action increase in frequency, while
343 very dissimilar actions decrease in frequency. With repeated reinforcement there is a process of

344 gradual refinement that homes in on the action that produces DA release. In the case of action
345 sequences, we observe a similar gradual refinement process whereby credit assignment for the
346 action sequence is accomplished by early refinement for the actions most temporally proximal to
347 reinforcement, followed by later refinement for the more temporally distal actions.
348 Previous synaptic and cellular studies[36,37] proposed that DA reinforcement may act
349 retrospectively to reinforce behavior. By utilizing the closed loop system, we rigorously tested
350 this prediction. Since retrospective reinforcement of behavior is not confined to the target action
351 alone, it facilitates credit assignment to a stimulation-producing action even when reinforcement
352 is delayed; stimulation-producing action pairs that tend to be performed closed together in time
353 were learned much faster than pairs that tended to be performed far apart in time. Intriguingly,
354 animals eventually learned to assign credit to distal stimulation-producing actions even in the
355 latter scenario. This is characterized by a gradual process whereby early on, the median time
356 interval between distal and proximal target actions decreased and the repertoire proximal to
357 reinforcement was preferentially refined to favor the performance of the proximal target action.
358 As the distal target action became significantly more likely to occur within second timescale
359 distance prior to reinforcement, retrospective reinforcement of the correct stimulation-producing
360 sequences became increasingly likely, resulting in whole behavioral refinement for the distal
361 target as well, hence increasing sequence performance (Fig. 5g).
362
363 It has been suggested that retrospective reinforcement of behavior is mediated by DA modulation
364 of an eligibility trace left by action potential-triggered synaptic plasticity[10]. Studies of DA action
365 at the striatal synaptic level[36,37] indicate that the timescale within which retrospective
366 reinforcement may occur is on the order of a few seconds, but the behavioral consequences have

remained elusive until now. Our behavioral findings are consistent with cellular studies in that behavior occurring within a few seconds leading into DA stimulation are reinforced. It is also noteworthy that distal T1 refinement in two action reinforcement occurs after the closest T1 to DA stimulation has become more probable within a few seconds of stimulation. The cutoff of retrospective reinforcement by phasic DA activities within a few seconds could explain the sudden increase in sessions required to reach criterion frequency amongst animals that were reinforced for action pairs with initially longer median time separations. Retrospective behavioral reinforcement may be mediated by DA modulation of Ca2+ influx left by earlier spiking activities. Ca2+ influx triggered by NMDA receptors would increase adenosine 3',5'-cyclic monophosphate at thin distal dendrites of medium spiny neurons, leading to transient and localized protein kinase A activity specifically within the retrospective time window, as regulated by high phosphodiesterase activity[14]. Similar actions have more similar and overlapping striatal neural ensemble activities[21]. Arrival of DA upon activation of action-specific ensembles may reinforce not only a specific action, but also similar actions. As striatal ensembles specific to actions are activated and a trial of eligibility traces is left temporally, DA arrival could set the stage for retrospective reinforcement of a spatially graded repertoire of actions within a few seconds, resulting in the observed behavioral learning patterns. Future studies testing these ideas would clarify how synaptic plasticity and cellular ensemble activities integrate to produce a dynamic refinement process, resulting in the behavioral principles for credit assignment revealed here.

END OF MAIN TEXT

390 **Methods**

391 **Animals**: All experiments were approved by the Portuguese DGAV and Champalimaud Centre
392 for the Unknown Ethical Committee and performed in accordance with European guidelines.
393 They were also performed according to National Institutes of Health (NIH) guidelines and
394 approved by the Institutional Animal Care and Use Committee of Columbia University. 3-5
395 months old DAT-Cre male mice in the C57/BL6J background[23] were used.

396

397 **Sample Sizes, randomization, and blinding.** For sample size, we applied a power of 0.8,
398 significance of $p<0.05$, and standard variation of 20% of the mean. We determined sample sizes
399 of 4-8 mice per group for different mean-based tests (matched pairs, 2 groups). No formal
400 method of randomization was used; littermates were equally divided among the groups being
401 compared. The experimenter was not blinded of the experimental groups. Optogenetic
402 manipulations were performed automatically via a computer algorithm and not manually by the
403 experimenter.

404

405 **Recombinant adeno-associated viral vectors, stereotaxic injections, and implants**. 750 nl of
406 rAAV.EF1a.DIO.hChR2(H134R).eYFP or rAAV.EF1a.DIO.eYFP (3-4 x 10^12 vg/ml, AAV5,
407 University of North Carolina Vector Core; 1-2 x 10^13 vg/ml, AAV1, Addgene, 27056-AAV1
408 and 20298-AAV1) were injected into each hemisphere of the VTA of 3-4 month old DAT-Cre
409 mice. For viral injections, the coordinates are AP - 3.52 mm, ML - +/- 0.35 mm, DV – 4.3 mm.
410 Injections were made at 0.2 Hz pulses. Each pulse injects 4.6 nl volume. Injected needles were
411 kept in place in the injection site for ~15 minutes before withdrawal. For each mouse, a dual
412 optic fiber cannula (200/240 μm diameter, 6 mm length, 0.7 mm center-to-center FLT, 0.22 NA;

413 Doric, DFC_200/240-0.22_6mm_DF0.7_FLT) was placed 200 µm above the injection site and
414 fixed to the skull. Next, a 4-position receptacle connector (Harwin Inc., M52-5000445) was fixed
415 anteriorly to the dual optic fiber cannula, with its posterior edge set at -0.6 mm. Skull implants
416 are then fixed with dental cement. A 4-position connector (Harwin Inc., M52-040023V0445)
417 with pins removed from one end was used to cap the receptacle connector.
418
419 For photometry experiments, 3-5 month old DAT-Cre males were used. The conditions used for
420 VTA injections and implants were as above. Additionally, 1 µl and 500 nl of AAV9-hSyn-
421 GRAB-rDA1m (2 x 10^13 vg/ml; Addgene, 140556-AAV9) were injected into the dorsal
422 striatum (AP 0.5 mm, ML +2.1 (right), DV 2.3 (from brain surface)) and ventral striatum (AP
423 1.15mm, ML +1.65 (right), DV 4.2 (from Bregma)) , respectively. For photometry fiber
424 implants, mono fiberoptic cannula were used (400/430 µm diameter, 4 mm length (dorsal
425 striatum) and 6 mm length (ventral striatum), 0.37 NA, 1.25 diameter ferrule, flat; Doric,
426 MFC_400/430-0.37_6mm_MF1.25_FLT (ventral striatum) and MFC_400/430-
427 0.37_4mm_MF1.25_FLT (dorsal striatum)). Implants were inserted at a 22 degrees angle. For
428 dorsal striatum implantation, the cannula entered the skull at AP 0.5 mm and ML 3.03 mm at 22-
429 degree angle. The angled implant penetrated the brain from its surface for 1.92 mm. For ventral
430 striatum implantation, the cannula entered the skull at AP 2.85 mm at 22 degrees angle, ML 1.65
431 mm. The angled implant penetrated the brain from its surface for 4.25 mm.
432
433 **WEAR motion sensor system.** The WEAR motion sensor family was developed by the
434 Champalimaud Hardware platform and Costa lab as a wired or wireless solution to obtain self-
435 centered 9-axis motion data based on 3-axis accelerometer, gyroscope, and magnetometer

(https://www.cf-hw.org/harp/wear). The wired version is a very small and extremely lightweight device (200mg) that can sample motion data up to 500 Hz and at the same time provide current up to 500mA that can be used to power LEDs for optogenetic experiments or stimulating electrodes. The wireless version is small and lightweight (~1.8g) and can sample motion data up to 200 Hz while having the ability to provide up to 50 mA that can be used to power LEDs for optogenetic experiments or stimulating electrodes. The battery of the wireless WEAR allows recordings up to 4 h at 200 Hz sampling rate and even more at lower sampling rates. These devices communicate with the computers through a base station based on the HARP design developed by the Champalimaud Hardware Platform, which can be accessed through a software GUI to easily change sensor parameters to best fit the experimental needs. The base stations have several important hardware features such as 2 digital inputs and outputs, an analog input, 2 outputs for camera triggering, and a clock sync input and output that provides hardware-based synchronization. The sensor can be started or stopped by software or pin. The WEAR motion sensor family and base station are all open source (repository at https://bitbucket.org/fchampalimaud/workspace/projects/HP). Moreover, the WEAR devices are compatible with the Bonsai visual reactive programming software (https://bonsai-rx.org/), also open source, and allow the integration and synchronization of the streams of data being collected using the WEAR sensor with other data sources such as cameras.

Taking these specs and features together, the WEAR allows researchers to acquire high-resolution motion data wirelessly and for long periods of time, without being computationally very demanding. The 9-dimensional motion data acquired through WEAR is simple to process, easy to connect to analysis software, which allowed the fast online behavior classification that was fundamental for the experiments described in this paper.

459

460 **Open field experiment.** One-month post-surgery, mice were habituated to head-mounted
461 equipment over 2 days. On day 1, an actual or mock wireless inertial sensor (~2.5 cm H x 1 cm L
462 x 0.5 cm W with ~ 2.5-3.0 cm antennae, ~1.8 g weight) glued to the 4-position connector
463 (Harwin Inc., M52-040023V0445) was attached to the implanted receptacle connector on the
464 skull cap. Individual mice roamed freely in the home cage for 1 hour. On day 2, an actual
465 wireless inertial sensor and mono fiberoptic patchcord (200/220 μm diameter, 0.22 NA; Doric
466 DFP_200/220/900-0.22_2m_DF1.0-2FC) was attached to the skull cap via a mating sleeve.
467 Patchcords were attached to 1x2 fiber-optic rotary joint (intensity division, 0.22 NA; Doric,
468 FRJ_1x2i_FC-2FC) and mice roam freely in home cage for 1 hour. On open field recording day,
469 sensor/patchcord habituated mice were anesthetized by isoflurane, attached to equipment,
470 subjected to calibration protocol described below, and individually placed in an open field box
471 inside a sound insulated chamber. The open field box is made of 410 x 400 mm grey opaque
472 acrylic walls and a 410 x 400 mm white matte acrylic base. Individual mice were allowed to
473 behave freely inside the box for 75 minutes. The wireless inertial sensor (~1.8 g in weight,
474 WEAR wireless sensor v1.1; Champalimaud Scientific Hardware Platform) conveys motion
475 information sampled at 200 hz (set on WEAR v1.3.2 software; Champalimaud Scientific
476 Hardware Platform) to a receiver base-station (Harp basestation v1.1 or v. 1.2, Assembly v0,
477 Harp v1.4, Firmware v1.5; Champalimaud Scientific Hardware Platform), which conveys the
478 information to the experimental computer running a Bonsai script (Bonsai[38] editor v2.3.1) to
479 capture and record motion data and video information. Video was captured with a camera (Flea3
480 FL3-U3-I3Y3M(17450451), Point Grey Research) coupled to a 1/2" format lens (NMV-6WA,
481 Navitar).

482

483 **Calibration.** To ensure sensor stability within sessions, several approaches were employed.
484 First, a coated mating sleeve was attached to the dual optic fiber cannula that sits immediately
485 posterior to the sensor. The sleeve was thickened with black tape to a desired outer diameter such
486 that it stabilized the sensor in the anterior-posterior direction. Second, the metal pins in the 4-
487 position connector glued to the sensor were thickened with solder to stabilize their fit inside the
488 receptacle connector in the skull cap. This protects against displacement in all directions. Third,
489 stretchable black tape was wound around the base of the attached sensor and sleeve-covered
490 cannula, further protecting against shifts in sensor positioning.

491

492 To control for possible variation in sensor positioning across sessions, a calibration approach was
493 developed. Wireless inertial sensor was attached to individual isoflurane-anesthetized mice and
494 the sensor was secured with the above strategies. Next, individual mice was placed in a custom-
495 made calibration rig. The essential element of the rig is a vertical stainless-steel pole suspended
496 above a stably secured table. In the setup used, the vertical pole was fixed to the horizontal edge
497 of a vertically reversed "L" shape, stainless steel post assembly mounted on a breadboard
498 (Thorlabs). The space between the lower end of the vertical pole and the table is enough for an
499 individual mouse to slide underneath. The lower end of the vertical pole is fixed to a custom-
500 made connector that resembles the connecting end of the fiberoptic patchcord. To perform
501 calibration, individual isoflurane-anesthetized mice was securely attached to the vertical pole via
502 a mating sleeve bridging the connection to the mouse's cannula implant. Next, replicate readings
503 of the immobilized inertial sensor were made on Bonsai. Next, mice were attached to the
504 experimental patchcord and allowed to recover in home cage for 20 minutes or until individual

505 mice are clearly recovered and behaviorally active. Individual mice were then placed in open-
506 field box for experimentation.
507
508 Calibration involves rotating all accelerometer and gyroscope readings from the inertial sensor
509 by a rotation matrix such that the final gravitational field vector of the stationary sensor, when
510 mounted on the mouse and fixed to the calibration rig, is in a universal frame of reference
511 whereby there is zero vertical tilt. In other words, the only non-zero acceleration is on the
512 universal z-axis (pointing down). To accomplish this, the accelerometer pitch and roll orientation
513 angles of the fixed stationary accelerometer were determined and then applied to calculate the
514 rotation matrix. The rotation matrix is multiplied by the sensor accelerometer and gyroscope
515 readings to remove the stationary vertical tilt from the sensor. To account for possible drift in
516 gyroscope baseline over time, a daily reading of stationary gyroscope baseline was made with a
517 mock cement skull cap attached to the sensor just before the start of each experimental day. The
518 baseline gyroscope readings were subtracted from all gyroscope values before the rotation matrix
519 is applied to sensor data.
520
521 **Action Selection**. After open field run in the grey-walled box, off-line behavioral clustering was
522 performed on calibrated sensor data. To identify the natural action repertoire of individual mice,
523 we quantified behavior using acceleration and gyroscope time series features in a similar fashion
524 as described previously[21]. For the ground truth analysis, we used: 1.) Gravitational acceleration
525 (GA) along the anterior-posterior (A-P) axis for the discrimination of postural changes - GAap.
526 2.) Raw sensor acceleration along the dorsal-ventral (D-V) axis to quantify movement

527 momentum – ACCdv. 3.) D-V axis of gyroscope to extract head head-body rotational
528 information – GYRdv. 4.) Total body acceleration to differentiate resting state from movement.
529
530 Total body acceleration (TotBA) was defined as:
531
532 $TotBA = sqrt(BAap^2 + BAml^2 + BAdv^2)$,
533
534 where BAap, ml and dv represent the body acceleration of the anterior-posterior, medio-lateral
535 and dorsal-ventral axis, respectively. We calculated each individual BA component by median-
536 filtering the raw acceleration signals followed by a fourth-order Butterworth high-pass (0.5Hz)
537 filter. For the gravitational acceleration (GA) axis, the BA components were subtracted from the
538 median filtered raw signal axis.
539
540 All four time series features were binned into non overlapping 300 ms long window segments[26].
541 The values of each bin and per feature were then discretized, using fixed thresholds, producing a
542 summary distribution of each segment. For GAap and ACCdv we used 10 equal size threshold
543 values, plus two added bins between the limits and infinity to capture an approximated
544 distribution of values within each window bin. For GYRdv we used 5 thresholds (0, ±50, ±100)
545 to discriminate left and right turns. For TotBA, a single threshold was used to separate moving
546 from resting. The threshold was kept constant for all experiments and was set to the average
547 value separating the bimodal distribution of logTotBA (natural logarithm of TotBA feature). For
548 each 300-ms window segment we get four resulting histograms, one for each feature. The feature

histograms were individually normalized to obtain probability distributions and used to calculate the pairwise similarities between segments.

We used the "earth mover's" (EM) distance as a measure of similarity[25]:

$$S = -(dEM/4)^2$$

where dEM is the sum of the normalized EM distances for the 4 features (GAap, ACCdv, GYRdv and TotBA) defined above. The bin normalizations constrain S values within the range [-1,0], specifically, -1 and 0 define the maximum dissimilarity and identity between the two probability distributions, respectively. Finally, to produce a continuous unbiased classification of behavioral states, the similarity measures were clustered using affinity propagation[20], with the preference parameter set to the minimal value of the similarity matrix; this particular value was used for its stable number of behavioral clusters within its range.

Using the behavioral clusters identified by affinity propagation clustering of the grey open field behavior[13] as a ground truth for the true identity of each 300 ms histogram, we were able to simulate and evaluate the precision with which the Earth Mover's Distance (EMD) metric[21,25] could be applied for cluster matching online. Notable difference between the EMD metric used here is the use of the 4 features mentioned above rather than the 3 features used previously[21], as well as the multiplication of the similarity score by -1 such that the range of possible scores from maximal identity to dissimilarity is 0 to 1, respectively. Although the EMD cluster matching outcome correlates strongly with affinity propagation clustering, some false positive and false

negatives may occur. Several filters were set to optimize cluster selection for reinforcement: 1.) We selected for clusters that show low false positive rate (<5.5%) and below the 60th percentile false positive rate amongst all clusters per animal. 2.) We selected against clusters with high false negative rates (> 90th percentile of clusters per animal). 3.) We selected against clusters that tend to be performed serially within a short time interval. We calculated the probability that a target cluster or its top 5 most similar clusters (determined by EMD score) would reappear 3-18 seconds after the first occurrence of the target cluster. Clusters that tend to be repeated either by itself or have a high probability of having similar clusters appear within this 15 second window (> 90th percentile for median and range of probabilities of cluster appearing in window) were removed from selection pool. 4.) We filtered against clusters whose matching by EMD would be more sensitive to anterior-posterior shifts of the inertial sensor (although we already protected against this possibility with the safeguards above) (> 90th percentile for percent deviation from original cluster matching after shifts of accelerometer reading in the anterior or posterior direction). For each cluster, percent deviation is calculated first by summing up the total absolute cluster matching changes from original cluster matching data in the anterior and posterior shifted datasets. Next, the sum of deviation in the two altered datasets is divided by two and then divided by the total of cluster calls from the original dataset, and multiplied by 100 to get percent deviation from original cluster matching result. 5.) We selected for clusters that show fully accelerating movement (cluster exemplar value of less than the maximum value of 1 in the body acceleration feature bin of histogram). To choose dissimilar clusters per animal, an algorithm was written filtering clusters of each animal's repertoire based on the feature histogram values of each cluster's representative, or exemplar. Thresholds were set along the GAap and GYRdv features to divide cluster exemplars based on the distribution of values within these feature

histograms. For each repertoire, all histogram values from all cluster exemplars are pooled to create a pooled histogram. The range of bins with non-zero values for each feature are identified. The algorithm then filters cluster exemplars in the repertoire for non-zero values in the high, medium, low, or high+low value bins. For example, action A identification occurs by selecting for a cluster exemplar with median counts falling in the high GAap and GYRdy value bins. action B would then be selected by filtering for an exemplar with median counts falling in the low GAap and GYRdy value bins. This results in actions that are highly dissimilar. For example, EMD similarity scores comparing action A to action B almost always, except for 1 ChR2-YFP animal, fall in the more dissimilar end of a distribution of scores created by comparing action A to all actions in each animal. Hereafter, clusters will be referred to as actions.

Closed-Loop Optogenetics. For close loop optogenetics, a computer running a Bonsai script captured and recorded wireless sensor motion data and video information as described above in grey-walled open-field experiment. Here, data is also streamed to a custom MATLAB code which analyzes action composition changes over the course of action reinforcement, we used the EMD metric[21] to label individual 300 ms motion histograms with an action ID. For each arriving 300-ms segment we calculate the EMD distance between each cluster exemplar (or representative) of the ground truth cluster library from the grey open field behavior recording. The motion features histogram is assigned to the action for which comparison with the exemplar gave the lowest EMD score (most similar to target) amongst all comparisons. Decision making for stimulation has a range of 35-55 ms time gap between action performance and sent decision for stimulation. To trigger optogenetics, a Multi-Pulse Width Modulation (PWM) generator (Harp Multi-PWM Generator hardware v1.1, Assembly v1, Harp v1.4, Firmware v1.1; Harp

618 Multi-PWM Generator software v2.1.0; Champalimaud Scientific Platform) converts each
619 decision to trigger laser into electrical signals for 15 light pulses of 10 ms pulse duration at 25
620 Hz, with each train of pulses occurring over 600 ms and at 25% duty cycle. The multi-PWM
621 signal is passed through a 12 V, 7.2 W amplifier (Champalimaud Scientific Platform) and fixed
622 frequency driver (Opto-electronic, MODA110-D4-30 (2001.320220)) to control the activities of
623 a 473 nm, blue low noise laser (Shanghai Dream Lasers Technology, Co, Ltd. SDL-473-200T),
624 which was sent through an acousto-optic modulator (Opto-electronic, MTS110-A3-V1S (1001 /
625 330433)). The laser component that is modulated is then reflected by a mirror and funneled to a
626 mono fiberoptic patchcord, which is then coupled to a commutator. The output laser is then
627 passed through a dual-optic fiber patchord and connected to the implant cannula. Power
628 adjustment out of the tip of patchcord was made so that ~5mW was emitted from each end of the
629 dual optic fiber cannula. To ensure common time stamps from different channels, a clock
630 synchronization device (Harp Clock Sync v1.0; Champalimaud Scientific Platform) was
631 performed between the basestation and multi-PWM device.
632
633 **Single action sequence selection.** Mice were placed in a white open field box for closed loop
634 reinforcement protocol. Individual mice were subjected to a single session of protocol each day,
635 with sessions following each other on consecutive days. The white open field box is made of
636 410 x 400 mm white matte acrylic walls and a 410 x 400 mm white matte acrylic base. To
637 acquire baseline behavior, individual mice were allowed to behave freely inside the box for 30
638 minutes on the first action A selection session. Closed loop reinforcement by blue laser
639 stimulation of VTA DA neurons were made available for 60 minutes. 90 minutes of closed loop
640 reinforcement were made available for individual mice during sessions 2 and 3. For session 4, an

641 extinction protocol was carried out comprising of 20-minute maintenance of reinforced behavior
642 with laser availability, followed by 60 minutes of extinction of reinforced behavior without laser
643 availability, followed by 20-minute re-acquisition of reinforced behavior with laser availability.
644 To select for action B, a repeat of the protocol described above for action A was performed
645 starting on the day following extinction protocol of action A. Upon completion of the
646 reinforcement and extinction protocols for action B, a contingency degradation protocol was
647 performed comprising of 20-minute maintenance of action B with laser availability, followed by
648 60 minutes of contingency degradation of reinforced behavior by triggering laser randomly,
649 followed by 40-minute re-acquisition of reinforced behavior with laser availability for action B
650 performance.
651
652 **Photometry experiment.** One-month post-surgery, mice were habituated to head-mounted
653 equipment for 2 days. On day 1, habituation was made to wireless inertial sensor as described
654 above. On day 2, a multi-fiber bundled patch cord (3 fiber bundle, 400/440 μm diameter for a
655 maximum of inner diameter at 900 μm, 0.37 NA, 3.5 m long, 1.25 mm fiber tip diameter, low-
656 autofluorescence; Doric, BBP(3)_400/440/900-0.37_3.5m_FCM-3xMF1.25_LAF) was attached
657 to individual mice in addition to the wireless sensor and optogenetic patchcord. Individual mice
658 were allowed to habituate to the equipment for 1 hour in its home cage. On photometry recording
659 day, mice were subjected to 30 frames per second photometry recording (Neurophotometrics),
660 with 75-150 μW 560 nm LED illuminating rDA1m, and equivalent closed loop optogenetic
661 parameters described above were used. To test for DA release in the context of closed loop
662 optogenetic setup, an average of 30 hits of blue light were delivered randomly within the span of
663 30 minutes. To evaluate DA release in the context of food reward, mice were placed on food

664 deprivation protocol and kept within 85% of original weight. Mice were placed in an operant
665 chamber with a nosepoke linked to a lick detector (PyControl). Each lick detection triggers
666 dispensing 2 µl 10% sucrose. Since animals tend to accidentally trigger lick detector at the
667 beginning of sessions, between 40-50 sucrose dispensing events were gathered per animal and
668 rDA1m activities associated with the last 35 rewards of the session were used for analysis.
669
670 **Two action sequence selection.** Two action sequence selection occurs as follows: after
671 sensor/patchcord habituation and grey open field behavior recording, offline behavioral
672 clustering and action filtering were performed as for single action selection. For each animal,
673 median time intervals between all possible pairs of actions during open field were calculated as
674 described above. Across animals, T1/T2 pairs with median T1→T2 interval values varying
675 between 2 and 10 seconds, and with the feature of going from a head down(T1) to a head up(T2)
676 movement, were chosen for reinforcement.
677
678 On the first reinforcement session, a 30-minute baseline was taken when laser stimulation was
679 not available for reinforcement. Laser became available for reinforcement in all subsequent
680 sessions until extinction experiment. During reinforcement periods, when closed-loop system
681 detects performance of the proximal action (T1) of interest, the algorithm enters a state where
682 laser is triggered upon performance of the distal action (T2), regardless of the amount of time
683 that has elapsed between the latest T1 and T2. On Session 1, 60 minutes of laser availability was
684 given while in all subsequent reinforcement sessions, 90 minutes of laser availability was given.
685

686 **Histology and Immunohistochemistry.** After behavioral sessions were completed, mice were
687 deeply anesthetized with isoflurane and perfused transcardially in PBS and then 4% PFA/PBS.
688 Dissected brains with skulls attached were perfused in 4% PFA in PBS at 4 degrees Celsius
689 overnight. The next day, brains were rinsed 3 times in PBS. Next, brain regions including VTA
690 and implants were sectioned by vibratome into 50 or 100 μm slices. Slices are then subjected to
691 immunohistochemistry using the reagents below. Standard immunohistochemistry protocols
692 were applied to stain for the following reagents - Rabbit anti-GFP 488 conjugate (1:1000;
693 Molecular Probes A21311). Mouse Anti-TH (1:5000; Immunostar Th 22941) with Goat Anti-
694 Mouse - IgG (H+L) Highly cross-adsorbed secondary antibody - Alexa Fluor647 (1:1000;
695 ThermoFisher, A-21236), DAPI (1:1000 of 20 mg/mL stock; Sigma, D9542).
696
697 **Imaging.** Zeiss Axio Imager M2 microscope was used to acquire brain section pictures. 10x tiled
698 images were taken through the relevant fluorescent channels. The M2 is equipped with a fast
699 Colibri.7 LED illumination for excitation of fluorophores. Images are captured with a high-
700 sensitivity monochromatic sCMOS camera (Hamamatsu Orca Flash 4.0 v2). The objective used
701 for the images is a ZEISS Plan-ApoChromat 10x/0.45, which allows to resolve up to 577 nm
702 when using a wavelength of observation of 520nm and it is fully corrected for chromatic and
703 spherical aberrations. Implant locations were determined using standard mouse atlas[39].
704
705 **Single action selection** analyses. For target action frequency analysis, we analyzed frequencies
706 within 25-minute windows at 4 time points: Baseline (before first reinforcement trigger), Early
707 (after first reinforcement trigger in Session 1 (action A) or 5 (action B)), Mid (after 2-minute
708 mark in Session 2 (action A) or 6 (action B)), Late (after 2-minute mark in Session 3 (action A)

709 or 7 (action B)). For 3D action repertoire plots, baseline normalized frequencies were plotted and
710 actions whose time series include NaN or Infinity values were discarded from the plot. (Plotted
711 actions: 509 of 514 actions, 15 ChR2YFP animals (action A); 427 of 443 actions, 13 ChR2YFP
712 animals (action B); 355 of 356 actions, 10 YFP animals (action A); 341 of 356 actions, 10 YFP
713 animals (action B)).
714
715 Three parameters were assessed for rapid behavioral adaptation following cumulative closed
716 loop reinforcements: latency between Target A triggered reinforcements, Target A frequency and
717 average behavioral similarity to Target A. To calculate the latency parameter, the average
718 latency between 10 consecutive Target A triggered reinforcements following a specified number
719 of cumulated reinforcements were taken and then normalized by the average latency taken over
720 the final 10 baseline Target A instances that in simulations would have triggered reinforcement.
721 To calculate the frequency parameter, the frequency of Target A triggered reinforcements over
722 the course of 1 minute following a specified number of cumulated reinforcements were taken and
723 then normalized by frequency of the final 10 baseline Target A instances that in simulations
724 would have triggered reinforcement. To calculate the behavioral similarity parameter, the
725 average behavioral similarity (EMD score) to Target A between 10 consecutive Target A
726 triggered reinforcement events following a specified number of cumulated reinforcements were
727 taken and then normalized by the corresponding value taken over the final 10 baseline Target A
728 instances that in simulations would have triggered reinforcement.
729
730 **rDA1m Fiber Photometry Analyses**. To evaluate DA release in the context of food reward, the
731 delta F/Fo signal was plotted for rDA1m signal aligned to lick detection/reward trigger. The

732 baseline Fo value was taken as the median rDA1m raw fluorescence signal of the 10 time points
733 (333.33 milliseconds) preceding the trigger event. To test whether DA release is triggered in the
734 context of the closed loop system, the activity of the rDA1m sensor was quantified. Delta F/Fo
735 was calculated by subtracting baseline value from each fluorescent rDA1m value of a
736 smoothened time series (smooth function, default moving average filter, MATLAB), and then
737 dividing the outcome by the baseline value. To account for control ChR2-independent effects,
738 the average delta F/Fo trace of ChR2-YFP animals were subtracted from the corresponding
739 average trace of YFP animals, giving the differential delta F/Fo used for the plots. The standard
740 deviation of ChR2-YFP minus YFP curves were obtained by taking the square root of the sum of
741 squared variances of ChR2-YFP and YFP delta F/Fo curves.
742
743 **Categorizing behavioral actions by temporal dynamics.** To categorize behavioral actions by
744 temporal dynamics, moving mean of action counts was used as input. Various window sizes
745 were examined; 2.5-minute windows moving at 300 ms steps were found suitable for analyses.
746 The baseline frequency (f0) was the average of 5 minutes of moving mean data preceding the
747 first reinforcement event. Early frequency rate (f1) was the average of 30 minutes moving means
748 immediately following the first reinforcement event. Mid- and Late frequency rates were taken
749 from Day 2 (f2) and Day3 (f3), respectively. f2 and f3 rates were calculated from the beginning
750 30 minutes period after moving windows has accumulated enough bins (2.5 minutes) following
751 the start of the session. Significant positive modulation above baseline was judged if in 500
752 consecutive moving windows (2.5 minutes period) in Early/Mid or Late stages the frequency rate
753 of all bins were greater than the 99th percentile bin of baseline frequency. Significant negative
754 modulation below baseline was judged if in 500 consecutive moving windows (2.5-minute

755 period) in Early/Mid or Late stages the frequency rate of all bins were less than or equal to the 5th
756 percentile bin of baseline frequency. Actions that showed both significantly positive and
757 negative modulation at Early/Mid or Late stages when compared to baseline were delegated to
758 positive modulation group. For figure plotting, time-course median frequencies of action
759 dynamic types were downsampled 10-fold. To investigate the relationship between target
760 similarity and frequency, two approaches were taken. To perform multiple comparison statistics,
761 actions were binned by their percentile ranking in terms of similarity to target (EMD). This is
762 because action distribution based on raw EMD binning was not even. Percentile binning allowed
763 for even distribution of actions amongst the groups. To examine the distribution of action
764 dynamic type frequencies in terms of target similarity, a binning by raw EMD score (0.5 score
765 binwidth) was used because this allowed for clear visualization of the relationship between target
766 similarity and frequency. Alternatively, percentile binning of EMD score was also used and gave
767 similar trends.
768
769 **Criterion for action dynamic types.** Action dynamics were grouped according as follows: 1.)
770 Increasing actions showed significant increase in f0 to f1/2 and f1 to f2/3 comparisons and
771 showed either significant increase or unchanged frequency in f1/2 to f3 comparisons. 2.)
772 Sustained actions showed significant increase in f0 to f1/2 comparisons, and unchanged
773 frequency in f1 to f2/3 and f1/2 to f3 comparisons. 3.) Transient actions showed significant
774 increase in f0 to f1/2 comparisons, and significant decrease in f1/2 to f3 comparisons. 4.)
775 Decreasing actions showed significant decrease in f0 to f1/2 and f0 to f3 comparisons. 5.) Other
776 actions were all remaining actions that did not fall in the above groups. In the main figure only
777 dynamic subtypes with more than 10 members are shown.

Extinction analyses.10 minutes portions from different time windows along the extinction protocols (Session 4 for action A and Session 8 for action B) were chosen. Early maintenance (M^1) starts from the first instance of target action performance in the session. Late maintenance (M^2) is the portion preceding the first performance of target upon extinction. Early extinction (E^1) begins at the first instance of target performance upon extinction. Late extinction (E^3) is the portion preceding the first performance of target upon re-acquisition. Mid extinction (E^2) begins at the midpoint between the starts of E^1 and E^3. Early re-acquisition (R^1) starts at the first performance of target upon re-acquisition condition. Late re-acquisition (R^2) is the final portion of the extinction protocol.

Action burstiness analysis. To evaluate action burstiness, or dispersion, we used Fano factor (variance/mean) as a measure. A survey of moving mean frequencies of reinforced actions across animals suggest that actions are more dispersed during the extinction phase, but the timescale with which this may occur is variable. To identify a suitable timescale to detect dispersion across reinforced actions, we screened a range of window sizes (600 ms to 5 minutes windows in 600 ms steps) with which to calculate moving window frequencies, and then calculate Fano factor in varying time segments. We chose a moving window of 15 seconds (50 x 300 ms action units) to construct moving mean frequencies. This window size consistently gave decreased Fano factor in baseline vs. maintenance session across animal, a result that would be expected as reinforcement led to stable target action performance.

Single action reinforcement, inter-target, and inter-action interval analyses. To quantity inter-target action intervals, the median amount of time that transpired between the start of successive target actions over the course of a time window was calculated. The time periods analyzed were: 1.) Baseline from the start of Day 1 (Sessions 1 and 5 for action A and B, respectively) until the first reinforcement event. 2-4.) Days 1 to 3 reinforcement. For reinforcement periods, behavior from the start of the first reinforcement event of that session until the end of session were analyzed. We considered the possibility that including the time interval between consecutive repeating of target actions (resulting in an inter-target action interval of 300 ms) would greatly affect the result. To test this, we removed values collected from consecutively repeating target actions. However, this did not affect result interpretations. Thus, we included intervals from consecutively repeating target actions in the presented analyses. For single action reinforcement, the median amount of time between the closest occurring action of interest and target action was calculated for both pre-target and post-target intervals.

Multinomial logistic regression predicting action dynamic types. To test whether intrinsic and baseline action properties are predictive of classifiable action dynamics during single action reinforcement from naïve state, two factors were considered. The factors are Earth Mover's Distance (EMD) similarity of action to target and median time interval of closest action of interest prior to target appearance at baseline condition.

To perform multinomial logistic regression, data from both dependent variables were log-transformed after addition of a constant value of 1. Transformed data were tested for collinearity

823 by examining scatter plots, Pearson's correlation coefficients, Variance Inflation Factors (VIF)
824 and condition indices. The two variables showed some correlation, but the coefficient value was
825 not above typical thresholds[40,41] and direct collinearity diagnostics did not show significant
826 collinearity (Pearson's correlation: $0.67 < 0.8$[40], VIFs: $1.82 < 5\text{-}10$[42], condition indices: $6.6 < 10\text{-}$
827 30[43]). Multinomial logistic regression was performed using MATLAB functions mnrfit and
828 mnrval. Non-Target A actions from all animals from reinforcement of action A were included
829 except those whose reinforcement dynamics were previously classified as "Other" types (n = 30
830 actions from a total of 514 actions, 15 ChR2-YFP animals). Decreasing dynamics type actions
831 were used as the reference group. Model accuracies were assessed using a 20-repeat, 10-fold
832 cross-validation approach for a total of 200 unique models for Real data, and 10,000 unique
833 models from 50 shuffled datasets.
834
835 To evaluate multinomial logistic regression, the deviance measure was used to judge model
836 fitting. Model performances were judged by area under precision-recall curve as this criterion is
837 suitable for imbalanced categories in the data[35]. A model containing both dependent variables
838 was found to outperform that of any single variable, even after consideration for penalties for an
839 extra factor (Akaike Information Criterion). The lack of significant collinearity between
840 dependent variables was supported by the stability of two relevant parameters, beta-coefficient
841 directions and significant p-values, across 200 cross-validation models and single- and double-
842 factor regression conditions (See Supplementary Information for tables).
843
844 **Dopamine retrospective window analysis.** To analyze whether DA reinforces actions proximal
845 to target, baseline rates of action transitions occurring close to reinforced action were examined.

846 First, a matrix tabulating 300 ms action counts from 2.4 seconds before to 2.4 seconds after each
847 theoretical target-triggered laser stimulation (600 ms in length) during baseline condition was
848 constructed. Next, all possible 600 ms action transitions (ex. X→Y) for each animal were then
849 counted using the above matrix, resulting in an action transition type (row) vs. time bin (column)
850 matrix where the counts of each action transition type occurring in specific 600 ms transition
851 windows (ex. X→Y) were recorded (sum across rows). This will be called the count matrix.
852 Next, the relative enrichment of each action transition type in a specific transition window
853 against all transition windows was calculated by dividing the action transition count matrix by
854 the total number of action transitions per type (probability across rows). Next, action transition
855 probability within a sliding 1.2 second transition window (containing a total of three action
856 transitions) relative to surrounding temporal environment (3.6 seconds) was derived by
857 subtracting the total number of action transitions per type within the surrounding 3.6 second
858 window from the total number of action transitions per type within the 1.2 second sliding
859 window of interest. This will be called the differential probability matrix. Next, action transition
860 types that showed greater than a threshold of 0.001 relative probability within sliding 1.2 second
861 windows of interest over the corresponding surrounding windows were filtered and kept for the
862 next step. Next, for each sliding 1.2 second window, the count matrix from above was analyzed
863 to select for action transition types that occurred between 2 to 6 times during the 30 minutes
864 baseline period (0.067 to 0.2 action transitions per minute). The count range was chosen to filter
865 out single events while selecting for action transitions with low initial frequencies over the
866 baseline period and analysis time range. Since the range of probabilities of specific action
867 transition types could vary greatly between different sliding 1.2 second windows, filtering as
868 above also balances the distribution of action transition probabilities amongst all action transition

869 types analyzed across sliding 1.2 second transition windows. The above process results in a list
870 of action transition types enriched for each sliding 1.2 second transition window, and baseline
871 normalized frequencies of these action transition types upon reinforcement in subsequent
872 sessions were calculated. Note that baseline normalized frequencies were calculated from all
873 occurrences of specific action transition types, regardless of their time distance in relationship to
874 target occurrence. Baseline normalized frequencies of individual action transition types were
875 averaged within animals and the means between animals are averaged to produce animal-
876 balanced results. Identical data trends and conclusions could be reached even if baseline
877 normalized frequencies of all action transitions were used for analyses.
878
879 **Two action sequence experiment analyses.** Two action sequence frequency was quantified in
880 terms of laser triggers per minutes. To assess learning across animals, the baseline frequency was
881 subtracted from frequencies of all reinforcement sessions. A criterion baseline subtracted
882 frequency of 3.2 triggers per minute was set after considering the range of baseline subtracted
883 frequencies observed in the open field and reinforcement sessions all animals. The criterion is set
884 such that it is > 20 % above the highest baseline-subtracted frequency value seen at open field
885 condition. The criterion point consistently falls above the open field frequencies of all animals
886 and marks the rising phase of all reinforcement frequency curves.
887
888 T1→T2 intervals were quantified as the time distance between the end of the latest distal action
889 (T1) and the end of the proximal action (T2) that triggers laser. T2→T1 intervals were quantified
890 as the time distance between the end of T2 that triggers laser and the end of the next closest T1.

To produce equivalent measures in open field and baseline conditions, laser trigger events were simulated by scanning across the data as if reinforcement was available.

Significance testing was performed on 14 of 15 ChR2-YFP animals that reached criterion frequency (ChR2-YFP Criterion). The lone animal that did not reach criterion frequency was removed because the T1→T2 median interval was still very high after session 10. This animal was subsequently subjected to single action reinforcement protocol to assess its ability to learn T1 and subsequently T2. Next, the animal was again subjected to T1→T2 reinforcement protocol. These results indicate that this animal was capable of action learning for both T1 and T2 separately, and for T1→T2 sequence after learning of each individual action.

Reinforcement sessions for the 14 ChR2-YFP animals that reached beyond criterion frequency continued until the T1→T2 interval has been decreased to below at least a median of 2 seconds. As YFP animals do not decrease the T1→T2 median interval over sessions, we stopped reinforcement at session 20.

Two action sequence extinction. Extinction session begins with a 25-minute maintenance period for two action-sequence reinforcement, followed by a 40-minute extinction period when laser was inactive, followed by a 25-minute re-acquisition period whereby reinforcement was made available again. To quantify performance for plotting, frequency was calculated over 5 minutes bins and then normalized to the last 5 minutes bin of the maintenance condition. For significant testing, raw frequencies were analyzed at the last 5 minutes of maintenance, extinction, and re-acquisition conditions.

914

915 **Two action sequence refinement.** To measure refinement for T1 and T2 in the two-action
916 sequence, actions that were uniquely related to one but not the other were identified. Actions
917 performed by each animal in their open field repertoires were ranked by their EMD similarity
918 scores to T1 or T2. The top-12 actions (within action repertoires ranging between 30-40 actions)
919 most similar to either T1 or T2 were identified. Actions common to both T1 and T2 in these lists
920 were removed, leaving actions uniquely similar to T1 or T2. We required at least 3 non-target
921 actions to be uniquely related to each of T1 and T2. One of the animals did not meet this
922 requirement, because less than 3 actions were uniquely similar to each of T1 and T2 when
923 considering the top-12 actions related to T1 or T2. For this animal, we relaxed the stringency by
924 considering actions that uniquely belong as the top-9 actions most similar to either T1 or T2. We
925 took the median target-normalized frequency of these uniquely similar actions to T1 or T2 as the
926 refinement index. A refinement index of above or around 1 indicates little to no refinement of
927 uniquely related actions to target. Refinement index below 1 indicates refinement relative to
928 target; the lower the score the more refinement. Refinement curves were smoothened using the
929 Savitzky-Golay filter to improve visualization of trends. To better compare the progress of
930 refinement between T1- and T2-related actions, refinement indices were scaled such that the
931 minimum value amongst all sessions for individual animals would be zero and target-normalized
932 median frequency of 1 would remain at a scaled value of 1.

933

934 **Relationship between T1→T2 interval and sessions to criterion frequency.** To describe the
935 trend in a T1→T2 interval vs. sessions to criterion frequency scatter plot, non-linear sigmoidal fit
936 was tested against a 4th order polynomial fit. A linear fit was also tested. Sigmoidal fitting gave

937 the best result. The same fitting was tested for T2 → T1 interval vs. sessions to criterion
938 frequency, but the fit was poor and midpoint was unstable. For the T1→T2 sigmoidal curve,
939 half-maximum was 2.59 sessions to criterion frequency and midpoint was 4.69 seconds of open
940 field median interval. The half-maximum value was used to divide ChR2-YFP animals into slow
941 (above half-max) and fast (below half-max) learners.
942
943 **Differential refinement analyses.** The difference in area between T1 and T2 scaled refinement
944 curves over sessions was used to assess the relative refinement status between T1 and T2 over
945 sequence learning. The difference in areas were summed up using the trapezoid method across
946 sessions until the session when both T1 and T2 has or had reached minimal scaled refinement.
947 Next, the relationship between open field median interval and average difference in area under
948 T1 – T2 refinement curves per session was tested. Linear regression proved most suitable for
949 fitting (Goodness-of-fit: $R2 = 0.66$). The fit for T1→T2 linear line was $y = 0.1893x – 0.7050$.
950 Slope was significantly non-zero ($p = 0.0004$). The same fitting was tested for T2 → T1 interval
951 vs. difference in area under T1 – T2 refinement curves per session ($y = 0.00736x + 0.1356$), but
952 the fit was poor, and goodness of fit was low (Goodness-of-fit: $R2 = 0.07$). The slope was not
953 significantly non-zero ($p = 0.7063$).
954
955 **Starting Point identification for evaluating progression of differential T1/T2 refinement.** To
956 more precisely examine whether proximal action (T2) refinement precedes that of distal action
957 (T1) in Slow Learners, it was important to consider refinement progression of T1 relative to T2.
958 To rule out any bias towards proximal refinement because of initial bias towards proximal T2
959 refinement, a specific session was chosen as a Starting Point for analysis for each animal. This

960 Starting Point is defined by an early session in which T1 and T2 were relatively similar in
961 refinement levels or when the distal action T1 was more refined than proximal T2. To identify
962 these Starting Points, a scan was made retrospective from the session for which the T1→T2 time
963 interval is close to final value (less than or equal to a median of 3 seconds). Using this approach,
964 we identified earlier sessions in which distal T1 refinement was equal to or greater than proximal
965 T2 (T2 – T1 refinement curve area less than or equal to 0). The latest such session was set as the
966 Starting Point for analysis. If at no point early in learning did an animal have a session where
967 proximal (T1) action is most refined relative to distal (T2) action, an early session of closest T1
968 and T2 refinement was used as the Starting Point. The initial T2-T1 refinement curve area
969 difference calculated from the Starting Point to next session was subtracted from all T2-T1 area
970 differences calculated in subsequent sessions. This value is called the Starting Point subtracted
971 refinement difference. This made it possible to clearly track the change in relative refinement of
972 distal(T1) vs. proximal(T2) actions over time (Values above zero indicate T2>T1 refinement,
973 and values below zero indicate T1>T2 refinement). To identify the Turning Points for each
974 animal, sessions carrying the local maximum value of the Starting Point subtracted refinement
975 difference were identified for each animal. To calculate Starting Point subtracted refinement,
976 scaled refinement values from sessions of interest were subtracted from that of the Starting Point
977 session defined above.
978
979 **Odds ratio analysis.** For odds ratio calculation, the total amount of open field → Turning Point
980 session (second of two consecutive sessions used to calculate the refinement difference at
981 Turning Point as mentioned above) and Turning Point → session of criterion frequency median
982 interval changes were summed up for T1→T2 and T2→T1 intervals, respectively. Next, the

983 proportion of total interval change stemming from the open field condition→Turning Point
984 period, and from Turning Point→session reaching criterion frequency period, were calculated.
985 Next, the proportion of open field→Turning Point interval change was divided by the proportion
986 of Turning Point → session reaching criterion frequency period interval change for T1→T2 and
987 T2→T1 interval types, respectively. This gives the odds ratio.
988
989 **T1 probability rank and refinement change across time bins from T2 trigger.** For every
990 actual or simulated trigger for T1→T2 performance, the first occurrences of every action before
991 or after T2 triggers were counted at specific 300 ms time bins for up to 6 seconds before and
992 after T2 trigger. This was done for the specific conditions of baseline, Starting Point, Turning
993 Point, session passing criterion frequency, and last session. The probability of an action
994 occurring at a specific 300 ms time bin was calculated for all actions in the repertoire, and the
995 values were used to determine probability rank in terms of percentiles (100 percentile is most
996 probable action relative to all actions at a specific 300 ms time bin). To assess total T1
997 probability rank change within 0.3-1.8 or 2.1-3.6 second time bins, the area under the curve was
998 determined and values were normalized by subtraction from each animal's corresponding
999 baseline values. Refinement change was calculated by first taking the median probability rank of
1000 actions most uniquely related to T1 at varying time distances before or after T2 trigger. This
1001 value is then normalized by T1 probability rank to arrive at a refinement index. The area under
1002 the curve was determined and values were normalized by subtraction from each animal's
1003 corresponding baseline values. Decreasing values from Starting Point indicate increasing
1004 refinement.
1005

Statistical Analysis:

Standard statistical analyses were performed on Prism (GraphPad Software, Inc.) and permutation/bootstrap analyses were performed on MATLAB (MathWorks Inc.). To determine appropriate tests for comparisons, datasets were assessed for normality using Anderson-Darling, D'Agostino & Pearson, Shapiro-Wilk and/or Kolmogorov-Smirnov tests whenever applicable. Datasets were also visualized for normality using QQ plots and assessed for equal variance by examining the Residual plot (Residuals vs. Predicted Y). Parametric or non-parametric tests were chosen based on the combination of these analyses. Data were transformed logarithmically (with or without addition of a constant prior to transformation) whenever it was appropriate to promote normality and equal variance. Unless specified, sphericity was not assumed, and Geisser-Greenhouse correction was applied in all ANOVA tests. The appropriate post hoc multiple comparisons tests were applied to compare between the means of specific conditions wherever applicable. Significance was set at alpha $= 0.05$. For bootstrap analysis, significance was determined by asking whether the original target action mean Fano factor was greater or less than the 95% confidence interval of the bootstrap distribution. Permutation test was applied in the comparisons between regression models because of the large sample size discrepancy between groups. Bonferroni p adjustment was used to account for multiple comparisons in this case. For detailed description of statistical procedures please refer to Supplementary Information.

Acknowledgements:

We thank V.Athalye for helpful discussions and manuscript feedback, A. Vaz and C. Carvalho for mouse colony management, members of the Costa laboratory for comments, and the help from the Scientific Hardware Platform, Histopathology Platform, Scientific Software Platform and Advanced Bioimaging & BioOptics Experimental Platform (member of the Portuguese Platform of Bioimaging (PPBI-POCI-01-0145-FEDER-022122) of the Champalimaud Institute, S. Mutlu, D. Bento, P. Carriço., P. Silva, J. Araujo for hardware/software assistance, I. Marcelo for assistance with multinomial logistic regression, N. Loureiro for assistance with inertial sensor calculations , M. Mendoça and C. Alcacer for help with open field constructions. M. Carey lab for sharing apparatus. S. Fusi for project feedback. This work was supported by Life Sciences Research Fellowship and NINDS K99/R00 Award (1K99NS112575) granted to J.C.Y.T and National Institute of Health funding (5U19NS104649) to R.M.C.

Author Contributions:

J.C.Y.T and R.M.C. designed the study, interpreted results and wrote the paper. J.C.Y.T. performed and analyzed experiments. J.C.Y.T, V.P., F.C. designed close loop optogenetic system. J.C.Y.T., F.C. and A.S. executed assembly of the closed loop optogenetic system. F.C. and A.S. designed and assembled software and hardware. A.S. designed and assembled wireless inertial sensor and hardware. A.K. contributed Earth Mover's Distance code and was involved in early conceptions of the closed loop system. J.A.d.S., F.C. and A.S. designed and assembled the WEAR system. R.M.C. supervised the project. All authors edited the paper.

Competing Interests: F.C. is the Director of Open Ephys Production Site.

Additional Information: Supplementary Information is available for this paper.

Code availability. MATLAB (MathWorks) codes used for data analysis are available from the corresponding author.

Data availability. Source Data are available from the corresponding author upon reasonable request.

Correspondence and requests for materials should be addressed to rc3031@columbia.edu

References

1. Schultz, W. Predictive reward signal of dopamine neurons. *J. Neurophysiol.* **80**, 1–27 (1998).
2. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
3. Glimcher, P. W. Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proc. Natl. Acad. Sci. U. S. A.* **108 Suppl 3**, 15647–15654 (2011).
4. Thorndike, E. L. *Animal intelligence: Experimental studies.* viii, 297 (Macmillan Press, 1911). doi:10.5962/bhl.title.55072.
5. Skinner, B. F. *The behavior of organisms: an experimental analysis.* 457 (Appleton-Century, 1938).
6. Redgrave, P. & Gurney, K. The short-latency dopamine signal: a role in discovering novel actions? *Nat. Rev. Neurosci.* **7**, 967–975 (2006).

7. Minsky, M. Steps toward Artificial Intelligence. *Proc. IRE* **49**, 8–30 (1961).
8. Hull, C. L. *Principles of behavior: an introduction to behavior theory.* x, 422 (Appleton-Century, 1943).
9. Sutton, R. S. *Reinforcement learning an introduction /. Adaptive computation and machine learning* (MIT Press, c1998.).
10. Izhikevich, E. M. Solving the Distal Reward Problem through Linkage of STDP and Dopamine Signaling. *Cereb. Cortex* **17**, 2443–2452 (2007).
11. Reynolds, J. N. J., Hyland, B. I. & Wickens, J. R. A cellular mechanism of reward-related learning. *Nature* **413**, 67–70 (2001).
12. Shindou, T., Shindou, M., Watanabe, S. & Wickens, J. A silent eligibility trace enables dopamine-dependent synaptic plasticity for reinforcement learning in the mouse striatum. *Eur. J. Neurosci.* **49**, 726–736 (2019).
13. Fisher, S. D. *et al.* Reinforcement determines the timing dependence of corticostriatal synaptic plasticity in vivo. *Nat. Commun.* **8**, 334 (2017).
14. Yagishita, S. *et al.* A critical time window for dopamine actions on the structural plasticity of dendritic spines. *Science* **345**, 1616–1620 (2014).
15. Jin, X., Tecuapetla, F. & Costa, R. M. Basal ganglia subcircuits distinctively encode the parsing and concatenation of action sequences. *Nat. Neurosci.* **17**, 423–430 (2014).
16. Cui, G. *et al.* Concurrent activation of striatal direct and indirect pathways during action initiation. *Nature* **494**, 238–242 (2013).
17. Jin, X. & Costa, R. M. Start/stop signals emerge in nigrostriatal circuits during sequence learning. *Nature* **466**, 457–462 (2010).

18. Tervo, D. G. R. *et al.* Behavioral Variability through Stochastic Choice and Its Gating by Anterior Cingulate Cortex. *Cell* **159**, 21–32 (2014).
19. Skinner, B. F. 'Superstition' in the pigeon. *J. Exp. Psychol.* **38**, 168–172 (1948).
20. Frey, B. J. & Dueck, D. Clustering by Passing Messages Between Data Points. *Science* **315**, 972–976 (2007).
21. Klaus, A. *et al.* The Spatiotemporal Organization of the Striatum Encodes Action Space. *Neuron* **95**, 1171-1180.e7 (2017).
22. Boyden, E. S., Zhang, F., Bamberg, E., Nagel, G. & Deisseroth, K. Millisecond-timescale, genetically targeted optical control of neural activity. *Nat. Neurosci.* **8**, 1263–1268 (2005).
23. Lammel, S. *et al.* Diversity of transgenic mouse models for selective targeting of midbrain dopamine neurons. *Neuron* **85**, 429–438 (2015).
24. Dueck, D. Affinity Propagation: Clustering Data by Passing Messages. in (2009).
25. Rubner, Y., Tomasi, C. & Guibas, L. J. The Earth Mover's Distance as a Metric for Image Retrieval. *Int. J. Comput. Vis.* **40**, 99–121 (2000).
26. Wiltschko, A. B. *et al.* Mapping Sub-Second Structure in Mouse Behavior. *Neuron* **88**, 1121–1135 (2015).
27. Phillips, A. G. & Fibiger, H. C. The role of dopamine in maintaining intracranial self-stimulation in the ventral tegmentum, nucleus accumbens, and medial prefrontal cortex. *Can. J. Psychol. Can. Psychol.* **32**, 58–66 (1978).
28. Corbett, D. & Wise, R. A. Intracranial self-stimulation in relation to the ascending dopaminergic systems of the midbrain: a moveable electrode mapping study. *Brain Res.* **185**, 1–15 (1980).

1119 29. Sun, F. *et al.* Next-generation GRAB sensors for monitoring dopaminergic activity in vivo.
1120 *Nat. Methods* **17**, 1156–1166 (2020).
1121 30. Beier, K. T. *et al.* Circuit Architecture of VTA Dopamine Neurons Revealed by Systematic
1122 Input-Output Mapping. *Cell* **162**, 622–634 (2015).
1123 31. Howe, M. W. & Dombeck, D. A. Rapid signalling in distinct dopaminergic axons during
1124 locomotion and reward. *Nature* **535**, 505–510 (2016).
1125 32. Schultz, W. Behavioral Theories and the Neurophysiology of Reward. *Annu. Rev. Psychol.*
1126 **57**, 87–115 (2006).
1127 33. Dickinson, A. The 28th Bartlett Memorial Lecture Causal learning: An associative analysis.
1128 *Q. J. Exp. Psychol. Sect. B* **54**, 3–25 (2001).
1129 34. Elsner, B. & Hommel, B. Contiguity and contingency in action-effect learning. *Psychol. Res.*
1130 **68**, 138–154 (2004).
1131 35. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot
1132 when evaluating binary classifiers on imbalanced datasets. *PloS One* **10**, e0118432 (2015).
1133 36. Yagishita, S. *et al.* A critical time window for dopamine actions on the structural plasticity of
1134 dendritic spines. *Science* **345**, 1616–1620 (2014).
1135 37. Shindou, T., Shindou, M., Watanabe, S. & Wickens, J. A silent eligibility trace enables
1136 dopamine-dependent synaptic plasticity for reinforcement learning in the mouse striatum.
1137 *Eur. J. Neurosci.* **49**, 726–736 (2019).
1138 38. Lopes, G. *et al.* Bonsai: an event-based framework for processing and controlling data
1139 streams. *Front. Neuroinformatics* **9**, 7 (2015).
1140 39. Paxinos, George & Franklin, K. B. J. The mouse brain in stereotaxic coordinates / George
1141 Paxinos, Keith B.J. Franklin. (2001).

40. Berry, W. D., Feldman, S. & Stanley Feldman, D. *Multiple regression in practice*. (Sage, 1985).
41. Kim, J. H. Multicollinearity and misleading statistical results. *Korean J. Anesthesiol.* **72**, 558–569 (2019).
42. Neter, J., Kutner, M. H., Nachtsheim, C. J., Wasserman, W., & others. Applied linear statistical models. (1996).
43. Belsley, D. A., Kuh, Edwin. & Welsch, R. E. *Regression diagnostics identifying influential data and sources of collinearity*. (Wiley, 2004).

Figures and Figure Legends:

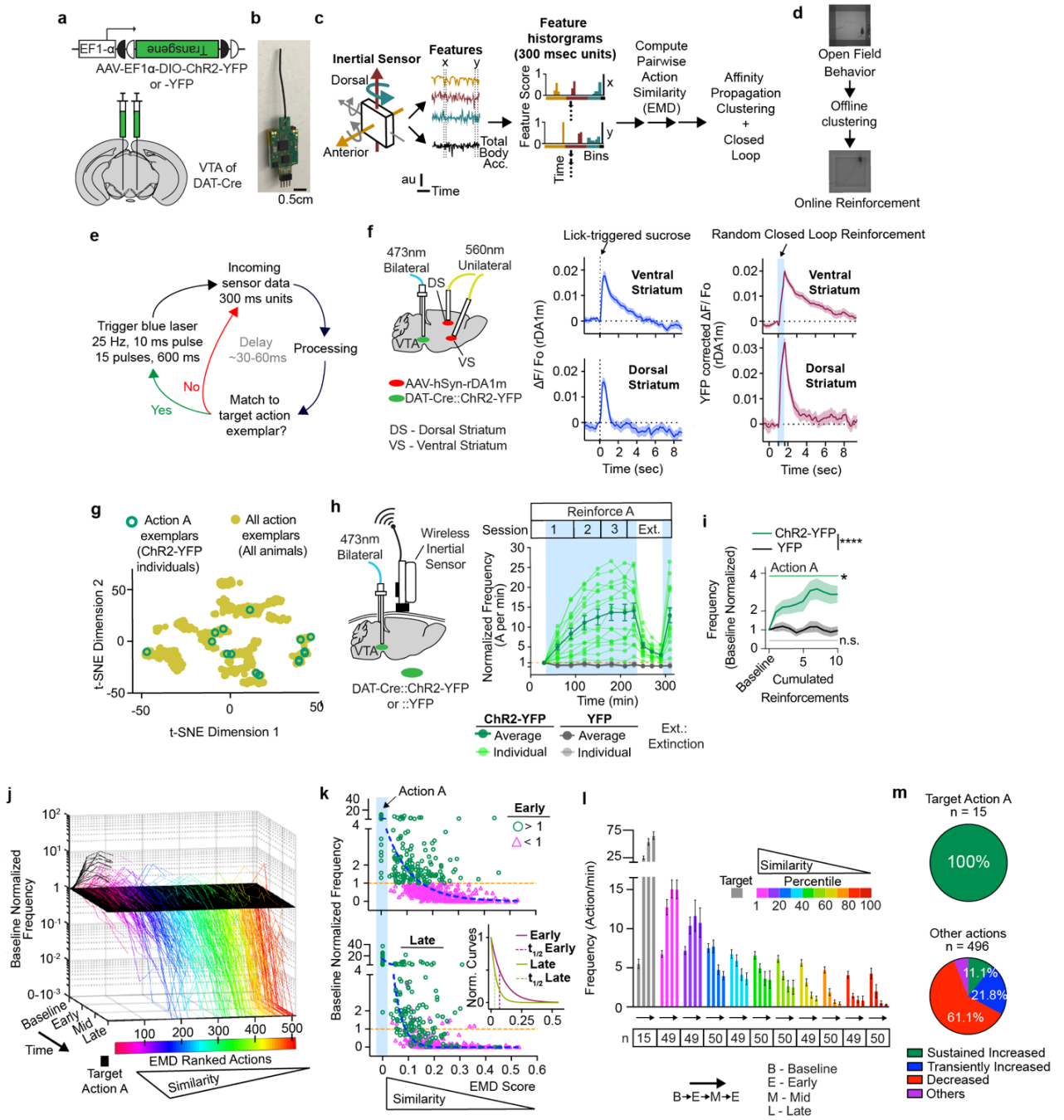

Fig. 1. **Learning of a single action from the naïve state as mediated by closed loop optogenetics. a**, Injection scheme. **b**, Wireless inertial sensor. **c**, Sensor data processing. **d**, Open field behavioral clustering and action reinforcement. **e,** Closed loop schematic. **f,** Dopamine release in dorsal and ventral striatum (n = 70 sucrose rewards, 2 ChR2-YFP mice; n = 66 and 65 random

stimulations, 2 ChR2-YFP and 2 YFP animals, respectively). Plots were mean, S.E.M. **g.** Action A exemplar locations in behavioral space. **h-m,** ChR2-dependent reinforcement of Action A (*n* = 15 ChR2-YFP animals (green). *n* = 10 YFP animals (grey)). Plots were mean, S.E.M. **h**, Left: Head-mount setup. Right: Light green/grey lines represent individual ChR2-YFP/YFP animals, respectively. **i,** Rapid increase in target action performance in response to close-loop reinforcements. Significant Time x Group Interactions (Supplementary Information). Plots were mean, S.E.M. **j,** Evolution of pooled behavior repertoire (n = 509 actions, ChR2-YFP mice) across learning. **k**, Early/Late cross-sectional views of (**j**) (Early: baseline normalized frequency >1, green circles, < 1, magenta triangles). Blue dashed lines - single phase log decay fits. Bottom inset graph shows Early/Late fitted lines normalized to 1 at EMD=0. **l,** Raw frequencies across learning and target similarity percentile groups. Plots were mean, S.E.M. Two-way mixed effects statistics in Supplementary Information. **m,** Pie chart summarizing distribution of actions according to their dynamics within reinforced Action A (left) or other actions (right). Asterisks: **** $p < 0.0001$. *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$. n.s. – not significant. See Supplementary Information for statistical/sample details.

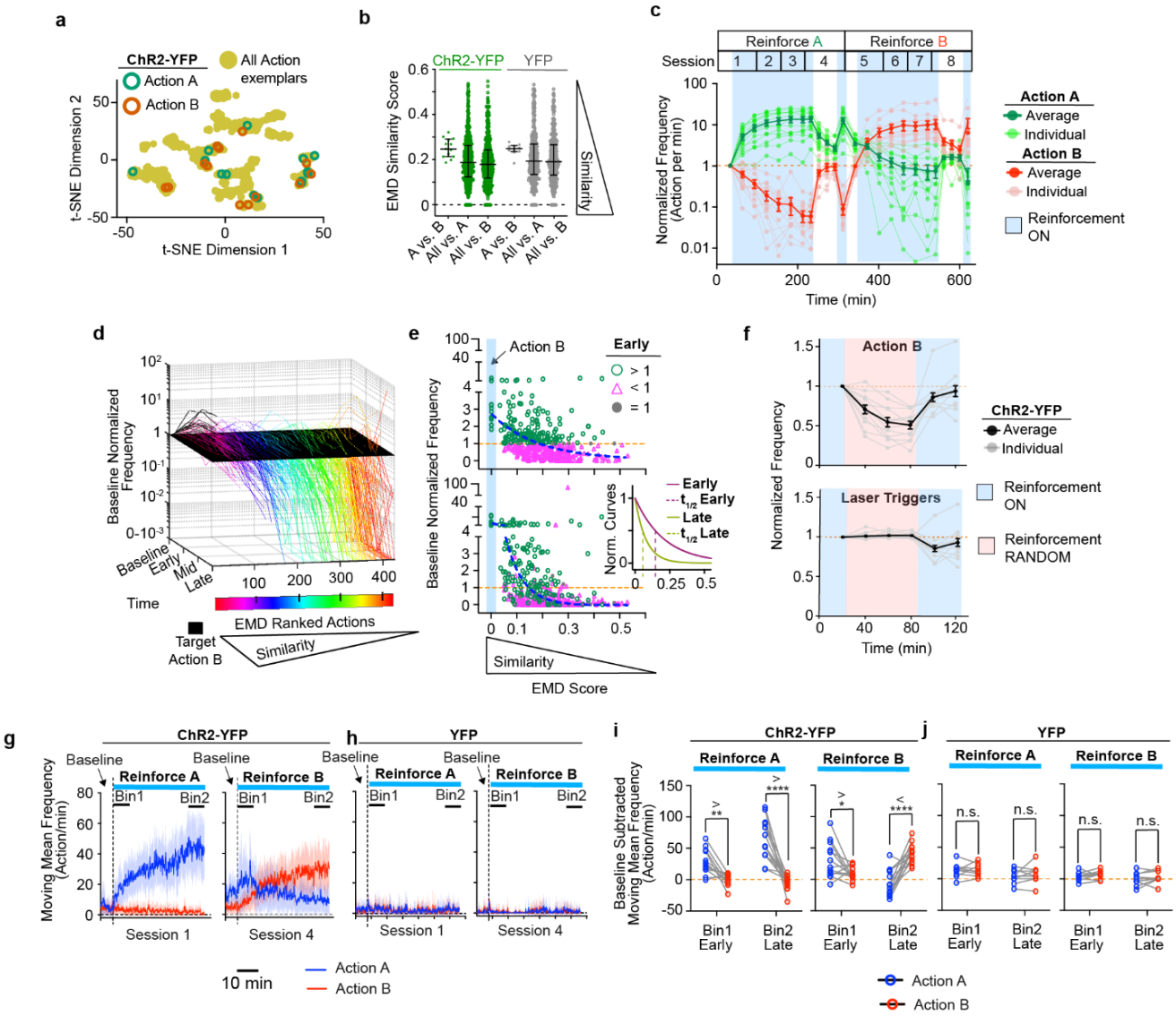

Figure 2. Transitioning from learned action to reinforcing new action. a-j, Animals reinforcing for Action A (n = 15 ChR2-YFP) to Action B (n = 13 of 15 ChR2-YFP). n = 10 YFP animals. **a.** Action A and B exemplar locations in behavioral space. **b,** Action similarity comparisons (A vs. B; n = 15/10, ChR2-YFP/YFP; All vs. A; n = 514/356, ChR2-YFP/YFP) or Action B (All vs. B; n = 443/356, ChR2YFP/YFP). Plot indicates median/interquartile range. **c,** Reinforcement for Action A and B in ChR2-YFP animals. Plot indicates mean/S.E.M. **d,** Evolution of pooled action repertoire (n = 427 ChR2-YFP actions) reinforced for Action B. **e,** Early/Late cross-sectional views of (**d**). Blue dashed lines indicate fitted decay curve. Bottom inset graph shows normalized Early/Late fitted

curves. **f,** Contingency degradation of Action B. Target random laser triggers frequencies (bottom) is based on initial Action B performance prior to contingency degradation. Plots indicate mean/S.E.M. **g-j,** Action A (blue) induced by reinforcement for Action B in experienced ChR2-YFP animals. **g-h,** Moving mean frequencies over reinforcement for Action A or B. Dashed, vertical line mark first reinforcement. Plots are mean/S.E.M (colored fill). Bin1/Bin2 are time bins for (**i-j**). **i-j,** Frequency measures within time bins noted in **(g,h).** Repeated measures two-way ANOVA reveal significant difference across time and actions A/B frequencies (not shown). Šidák's post hoc comparisons. Asterisks except in (**h**): **** $p < 0.0001$. ** $p < 0.01$. * $p < 0.05$. n.s. – not significant. See Supplementary Information for statistical/sample details.

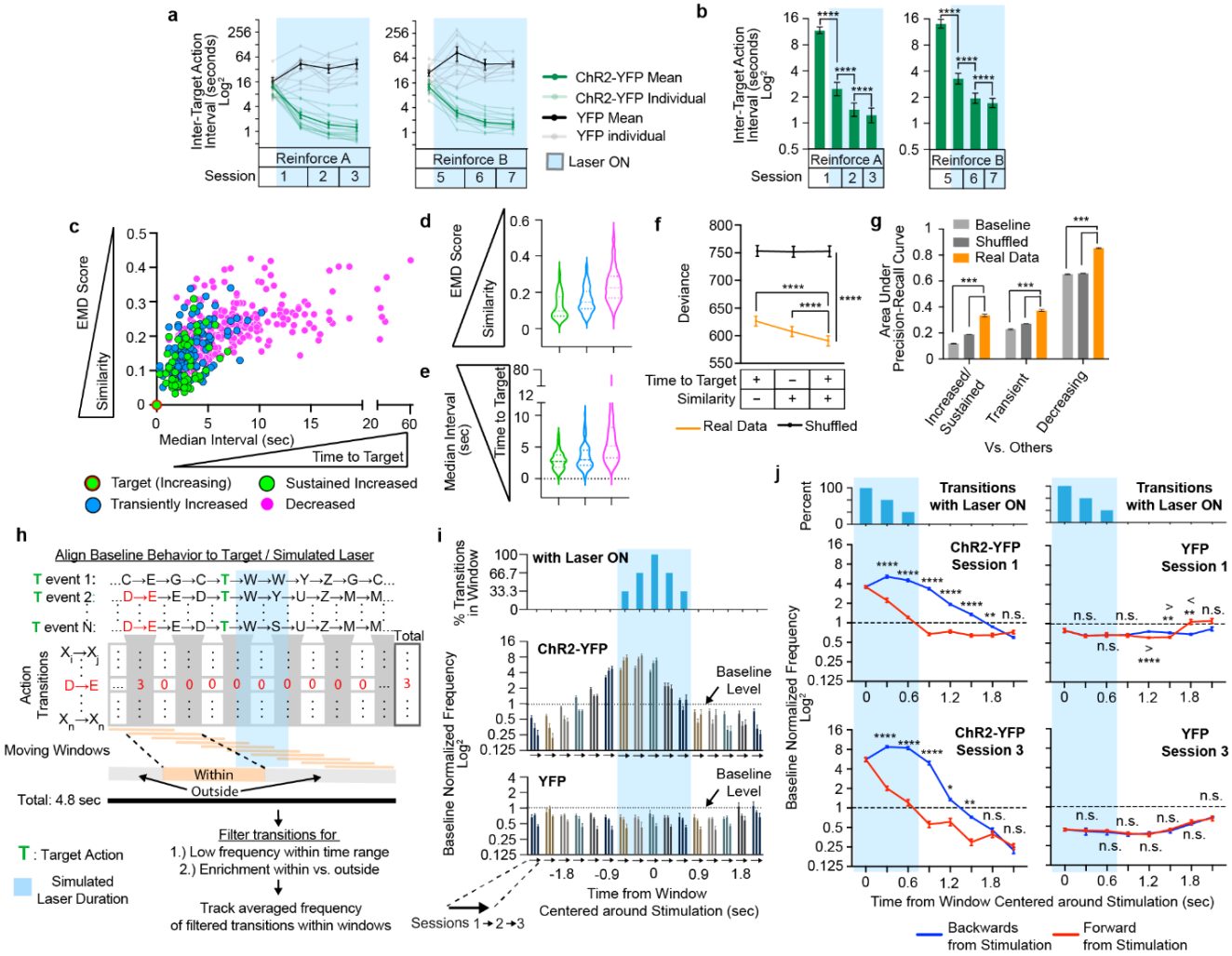

Figure 3. Dopamine mediates retrospective reinforcement of freely moving behavior. a-b, ChR2-dependent reinforcement decrease inter-action intervals for Action A (n = 15 ChR2-YFP) and B (n = 13 of 15 ChR2-YFP). n = 10 YFP animals. Plots are mean/S.E.M (**a-b**). Significant difference across time and ChR2-YFP/YFP (Mixed Effect Model. Action A: $F(3,69) = 72.26$, $p < 0.0001$. Action B: $F(3,62) = 33.78$, $p < 0.0001$.) **b**, Post-hoc Tukey's multiple comparisons of (**a**). **c-d**, Distribution of action dynamic types (n = 464 actions, 15 ChR2-YFP animals) according to target similarity (**c,d**), median time to target (**c,e**). **d-e**, Violin plots show median/quartiles. Two-tailed permutation tests with Bonferroni-adjusted p-values. **f-g**, Multinomial logistic regression of all factor combinations in Real data (200 models) versus Shuffled data (10,000 models). **f.** Groups differ across combinations (repeated measures, two-way ANOVA. $F(2,30594) = 518.2$, $p <$

0.0001.). Post-hoc Dunnett multiple comparisons. Plots are mean/std. **g**, Performance of double-factor regression model measured with area under the precision-recall curves (AUPRC). Two-tailed permutation test with Bonferroni-adjusted p-value. Plots are mean/S.E.M. **h**, Identifying moving window-enriched action transitions. **i.** ChR2-dependent reinforcement for Action A increases action transitions prior to and within stimulation window. Plots indicate mean/S.E.M. **j**, Quantification of (**i**). Significant difference across time and Retrospective/Forward reinforcement directions (Mixed Effect Modeling. ChR2-YFP Session1: $F(6,168) = 114.8$, $p < 0.0001$. ChR2-YFP Session 3: $F(6,168) = 46.62$, $p < 0.0001$, YFP Session1: $F(6,108) = 10.52$, $p < 0.0001$. YFP Session 3: $F(6,168) = 0.8992$, $p = 0.4984$). Post-hoc Šidák multiple comparisons. **** $p < 0.0001$, *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, n.s. – not significant. See Supplementary Information for statistical/sample details.

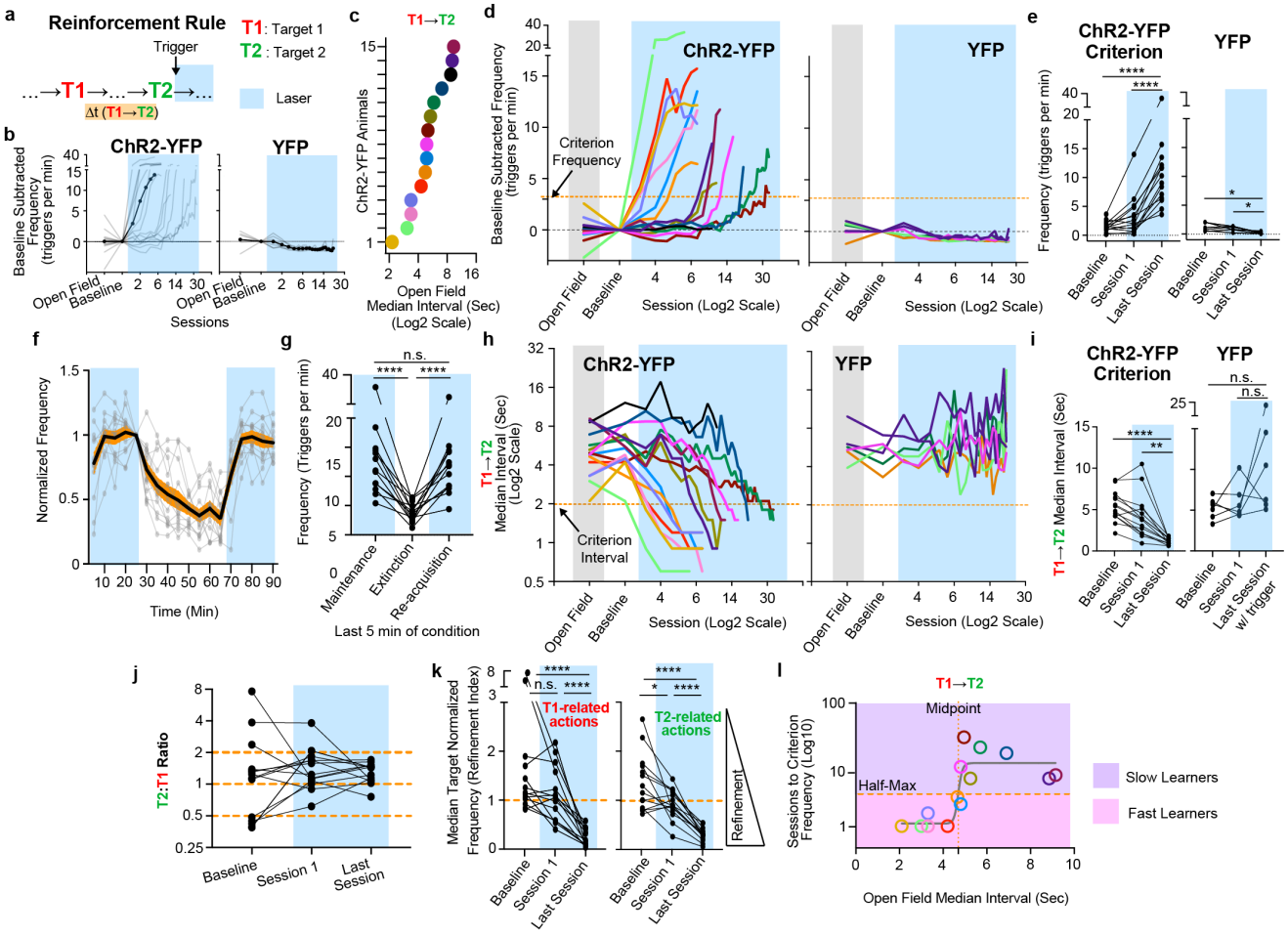

Figure 4. Relationship between pre-reinforcement inter-action intervals and learning of a two-action sequence. **a**, Schema. **b-l,** *n* = 15 (**b,d,h**) or 14 (**e-g,i-l**) ChR2-YFP, 6 YFP animals. Repeated measures one-way ANOVA, post hoc Šidák tests applied in (**e**,**g**,**i**,**k**). Plots of individuals in (**d-e**). **b**, ChR2-dependent increase in T1→T2 triggers (no laser during open field / baseline). **c**, Open field inter-action intervals of T1/T2 pairs chosen. Same color codes in (**d,h**). **d**, Individual learning curves labeled by color codes in (**c**). **e**, Frequency changes over conditions ($F(1.911,24.85)=51.02$, $p<0.0001$). **f-g**, Extinction of T1→T2 sequence (ChR2-YFP). **f**, Plot shows mean(black)/S.E.M.(orange fill)/individuals(grey). **g,** Frequency changes over extinction conditions ($F(1.073, 12.87) = 52.96$, $p<0.0001$). **h-i**, ChR2-dependent decrease in T1→T2 intervals. ($F(1.377, 17.90) = 35.95$, $p<0.0001$) (**i**). **j**, T2:T1 frequency ratios (ChR2-YFP) **k**, Target refinement shown by median target normalized frequencies of related actions. (T1: $F(1.237, 16.08) = 43.38$. T2:

$F(1.171, 15.22) = 48.74$. Both $p<0.0001$). Individual color code as in (c,g). **l**, Sigmoidal relationship between open field T1→T2 interval and sessions to criterion frequency.

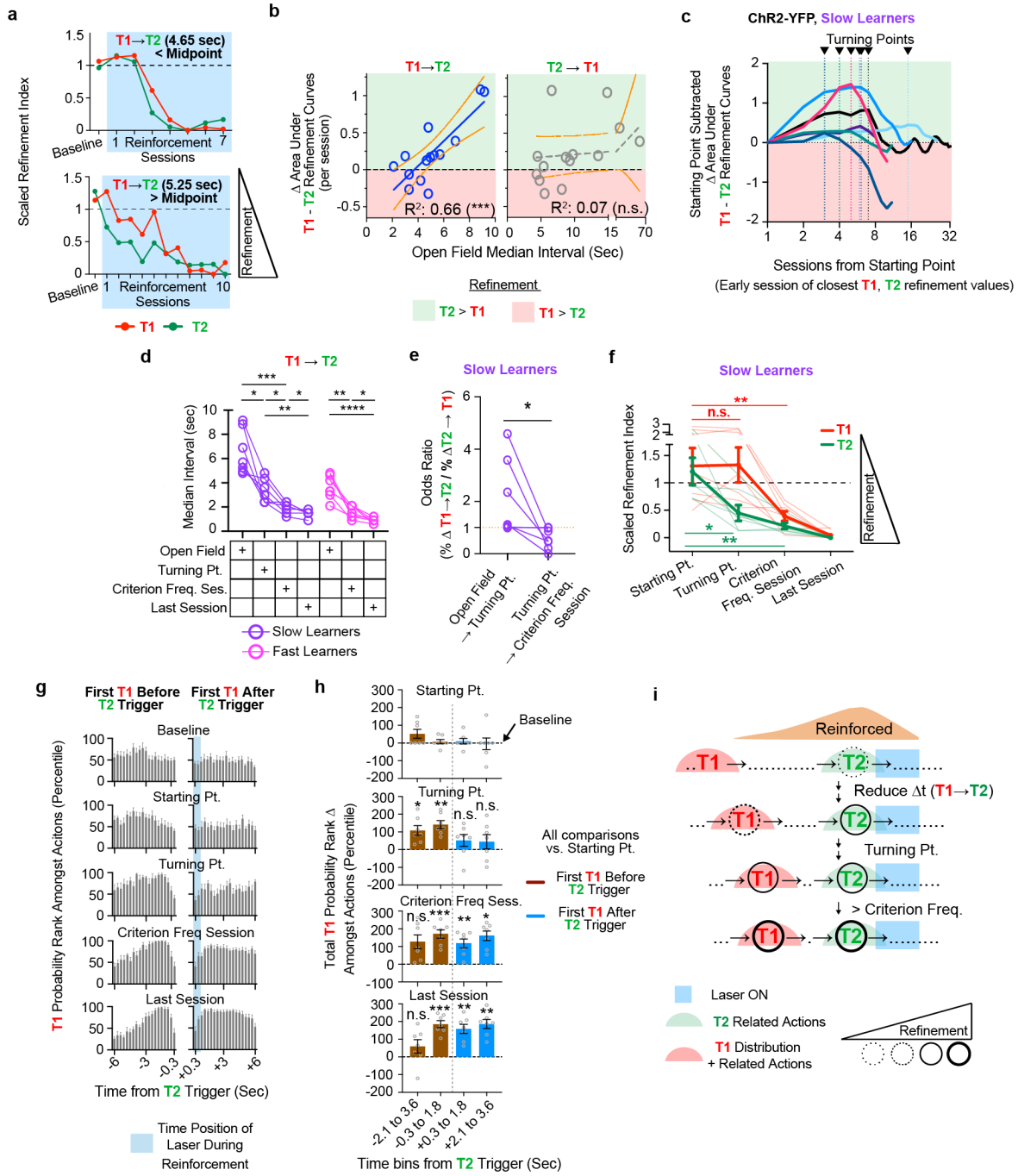

Figure 5. Behavioral process underlying learning of a two action sequence. *n* =14 ChR2-YFP (7 Slow Learners)**. a**, T1/T2 refinements in two ChR2-YFP individuals. **b**, Linear relationship between initial T1→T2 interval and differential T1-T2 refinement. Non-zero slope significance: T1→T2, p = 0.0004, T2→T1, p = 0.7063. **c**, Progression of differential T1-T2 refinement from Starting Point in Individual Slow Learners. **d,** T1→T2 interval significantly decreased by Turning Point in Slow

Learners. Repeated-measures 2-way ANOVA. Post hoc Tukey's test. **e,** Odds ratio of T1→T2 / T2→T1 interval changes. Paired Wilcoxon test ($p = 0.0312$, $n = 7$ animals). **f,** Preferential refinement of T2 relative to T1 by Turning Point in Slow Learners. Raw scaled refinement indices. Repeated measures, mixed effects model. Significant main effects. Time ($F(2.184, 26.20) = 54.21$, $p < 0.0001$). Post-hoc Šidák test. **g,** First occurrences of T1 before (left) and after (right) T2 triggers across learning stages. **h**, Quantification of pooled time bins from (**g**). Repeated measures, 2-way ANOVA for learning stage vs. rank change. First T1 Before and After T2 Trigger groups differ across learning stage and total T1 rank change. (Proximal bins (0.3-1.8 sec): $F(3,36) = 3.126$. $p=0.0376$. Distal bins (2.1 to 3.6 sec): $F(3,36) = 7.701$. $p<0.001$). Post-hoc Šidák relative to Starting Point values. **g,** Model for learning initially distantly separated T1→T2 sequences. Time not drawn to scale. **** $p < 0.0001$. *** $p < 0.001$. ** $p < 0.01$. * $p < 0.05$. n.s. – not significant. All bar plots indicate mean +/- S.E.M. See Supplementary Information for statistical/sample details.